Analysis of Expressed Sequence Tags Mapping to the Critical Region of the 5q- Syndrome

Thesis

Analysis of Expressed Sequence Tags Mapping to the Critical Region of the 5q- Syndrome

Amanda J. Strickson

A thesis submitted in partial fulfilment of the requirements of the Open University for the degree of Doctor of Philosophy

Leukaemia Research Fund Molecular Haematology Unit at the Nuffield Department of Clinical Laboratory Sciences, University of Oxford

February 2002

THE OPEN UNIVERSITY

RESEARCH SCHOOL

Research Degrees in Sponsoring Establishments

Library Authorisation

Please return this form to the Research School, The Open University, Walton Hall, Milton Keynes, MK7 6AA with the two bound copies of the thesis to be deposited with the University Library. All candidates should complete parts one and two of the form. Part three only applies to PhD candidates.

Part One: Candidate Details

Name: AMANDA JANE STRICKSON PI: R069568X

Degree: PhD Sponsoring Establishment: LEUKAEMIA RESEARCH F

Thesis title: ANALYSIS OF EXPRESSED SEQUENCE TAGS

MAPPING TO THE CRITICAL REGION OF THE 5q- SYNDRO

Part Two: Open University Library Authorisation

I confirm that I am willing for my thesis to be made available to readers by the Open University Library, and that it may be photocopied, subject to the discretion of the Librarian.

Signed: A. J. Strickson Date: 5 September 20

Part Three: British Library Authorisation [PhD candidates only]

If you want a copy of your PhD thesis to be available on loan to the British Library Thesis Service as and when it is requested, you must sign a British Library Doctoral Thesis Agreement Form. Please return it to the Research School with this form. The British Library will publicise the details of your thesis and may request a copy on loan from the University Library. Information on the presentation of the thesis is given in the Agreement Form.

Please note the British Library have requested that theses should be printed on one side only to enable them to produce a clear microfilm. The Open University Library sends the fully bound copy of theses to the British Library.

The University has agreed that your participation in the British Library Thesis Service should be voluntary. Please tick either (a) or (b) to indicate your intentions.

☑ I am willing for the Open University to loan the British Library a copy of my thesis.
A signed Agreement Form is attached

☐ I do not wish the Open University to loan the British Library a copy of my thesis.

Signed: A. J. Strickson Date: 5 September 20

Acknowledgements

The work described in this thesis was carried out at the Leukaemia Research Fund Molecular Haematology Unit, in the Nuffield Department of Clinical Laboratory Sciences, University of Oxford, under the supervision of Professor J.S. Wainscoat, Dr J. Boultwood, and Dr C.Fidler.

I would like to thank the Human Genome Mapping Project Resource Centre in Cambridge and the Resource Centre of the German Human Genome Project at the Max-Planck-Institute for molecular genetics for providing the cDNA filters and I.M.A.G.E. cDNA clones for this study.

I am particularly grateful to Jim Wainscoat, Jackie Boultwood and Carrie Fidler for their help and guidance through this project. I also thank them for their advice and constructive criticisms in the writing of this thesis.

Many thanks to the "IT" guys, Kingsley Micklem and Andrew Graham for their help with scanning and annotating images, and printing the thesis. Finally, I wish to thank Keith, Carrie and my parents for their love and moral support.

This work was supported by the Leukaemia Research Fund.

Abstract

Analysis of Expressed Sequence Tags Mapping to the Critical Region of the 5q- Syndrome

Amanda J. Strickson

Doctor of Philosophy

February 2002

The 5q- syndrome is a myelodysplastic syndrome characterised by a macrocytic anaemia, hypolobulated megakaryocytes, a low risk of transformation to AML, and a 5q- chromosome as the sole karyotypic abnormality. The approximate 5Mb critical region of gene loss of the 5q- syndrome has been defined in two patients with the 5q- syndrome at 5q31-q33, flanked by the genes for *FGF1* and *IL12ß*.

The frequent loss of genetic material from the long arm of chromosome 5 in association with a malignancy has led to the hypothesis that, by analogy with other malignancies characterised by genetic loss, the 5q- syndrome is caused by loss of function of a gene with tumour suppressor activity.

A transcript map of the 5q- syndrome critical region was generated with the aim of identifying the putative tumour suppressor gene associated with this disease. The expressed sequence tag (EST) database, db(EST) was used to isolate novel coding sequences mapping to the critical region of gene loss. Ten novel coding sequences (*C5orf4*, AF010242, AF156165, Cdy-17a06, Bda-87b11 195312, 4885953/143772, 120101, 195971, and 199067) were localised to the YAC contig spanning the critical region at 5q31-q33. The ten cDNA clones were sequenced, and overlapping clones were identified and sequenced in order to generate

complete or partial coding sequences. This included the cloning of novel gene, *C5orf4*, and the identification of the human synaptopodin and dynactin *p62* genes. In addition, the human homologues of the *Drosophila melanogaster RMSA-1* and *Saccharomyces cerevisiae CDC60* genes, and two known human genes (*PP2A* and *HAH1*) were localised to the critical region. Expression in human peripheral blood leukocytes and $CD34^+$ progenitor cells was investigated for each known and novel gene. Genomic localisation, expression patterns and predicted function would suggest these known and novel genes represent putative tumour suppressor genes.

Mutation studies were carried out on six known, and two novel candidate genes mapping to the narrowed 1.5Mb critical region of gene loss at 5q31.3-q32. No mutations were found in the coding regions/exons of these genes, suggesting they are not involved in the pathogenesis of the 5q- syndrome.

Contents

Acknowledgements		**2**
Abstract		**3**
Contents		**5**
List of tables and figures		**6**
Chapter 1	**Introduction**	**10**
Chapter 2	**Materials and methods**	**27**
Chapter 3	**Identification, localisation, cloning, and mutation analysis of novel gene *C5orf4***	**71**
Chapter 4	**Identification, localisation and analysis of novel cDNAs mapping to the critical region of the 5q- syndrome**	**105**
Chapter 5	**Analysis of species homologous ESTs mapping to the critical region of the 5q- syndrome**	**148**
Chapter 6	**Molecular analysis of the *SPARC*, *HAH1*, and *Annexin* VI genes**	**176**
Chapter 7	**Mutation analysis on five 5q- syndrome candidate genes by Denaturing High-Performance Liquid Chromatography (DHPLC)**	**208**
Chapter 8	**Conclusion**	**264**
Publications		**272**
References		**273**
Appendix	**Stock solutions and buffers**	**304**

List of tables and figures

Chapter 1

Figure 1.1	Ideogram of chromosome 5 showing some of the genes localised to 5q31-q33	19

Chapter 3

Table 3.1	MTN blots used in Northern analysis containing a variety of human tissues	78
Table 3.2	3' RACE PCR primer conditions for *C5orf4*	82
Figure 3.1	Northern blot analysis of I.M.A.G.E. cDNA clone 469867	88
Table 3.3	Overlapping I.M.A.G.E. cDNA clones identified from db(EST) homology searches from I.M.A.G.E. cDNA clone 469867	91
Table 3.4	Clones identified from screening the I.M.A.G.E. cDNA clone collection library with probe 209846	94
Figure 3.2	Transcription map of the critical region of the 5q- syndrome	96
Figure 3.3	Nucleotide and predicted amino acid sequence of *C5orf4*	98
Figure 3.4	Mutation analysis by cycle sequencing of patient 3 and *C5orf4*	100

Chapter 4

Table 4.1	RACE PCR primer conditions for novel cDNAs A3B02, 43911, and 199067	119
Table 4.2	YAC localisation PCR conditions for novel cDNAs 43911, 195312, 195971, 120101, and 199067	121
Table 4.3	ESTs identified from the collaboration with Genethon, and the Human Chromosome 5 GeneMap'98	123
Table 4.4	ESTs identified from the Human Chromosome 5 GeneMap'99	124
Figure 4.1	Gene dosage analysis of novel cDNA AF156165	125
Figure 4.2	Data map of novel ESTs identified from Genethon, the Human Chromosome 5 Genemaps'98 and '99, and the UniGene set	126
Table 4.5	Expression patterns and transcript sizes of novel cDNAs mapping to the critical region of the 5q- syndrome, identified from Genethon and the Human Chromosome 5 GeneMap'98	128

Table 4.6	Expression patterns and transcript sizes of novel cDNAs mapping to the critical region of the 5q- syndrome, identified from the Human Chromosome 5 GeneMap'99	129
Figure 4.3	Northern blot analysis of I.M.A.G.E. cDNA A3B02, and I.M.A.G.E. cDNA 240080	130
Figure 4.4	Hybridisation of high-density gridded, foetal brain cDNA library filters, with cDNA probes Cdy-17a06 and Bda-87b11	131
Figure 4.5	Dot blot analysis following hybridisation with cDNA probe Cdy-17a06	133
Figure 4.6	Southern blot analysis of I.M.A.G.E. cDNA clone A3B02	135
Figure 4.7	3' RACE PCR analysis of I.M.A.G.E. cDNA clone A3B02	137
Chapter 5		
Table 5.1	ESTs identified from the Human Chromosome 5 GeneMap and UniGene set	162
Figure 5.1	Gene dosage analysis of I.M.A.G.E. cDNA clone 194016	164
Figure 5.2	Northern blot analysis of I.M.A.G.E. cDNA clone 33583 (*CDC60*)	165
Figure 5.3	BlastX analysis of I.M.A.G.E. cDNA clone 33583 and the yeast cell division cycle gene *CDC60*	166
Figure 5.4	Sequence alignment of I.M.A.G.E. cDNA clone 145513 and the *Homo sapiens* chromosomal protein mRNA (the human homologue of the *Drosophila RMSA-1* gene)	167
Figure 5.5	Sequence alignment of I.M.A.G.E. cDNA clone 33583 and the *Homo sapiens* mRNA for leucyl tRNA synthetase (human homologue of the yeast *CDC60* gene)	169
Chapter 6		
Figure 6.1	Ideogram of chromosome 5 illustrating the critical region of the 5q- syndrome and the position of the *SPARC* gene	182
Table 6.1	PCR primer conditions for localisation of the *SPARC*, *HAH1*, and *annexin* VI genes to the YAC contig	188
Table 6.2	PCR primer conditions for $CD34^+$ expression analysis in the *SPARC*, *HAH1*, and *annexin* VI genes	188
Table 6.3	*SPARC* exon PCR conditions	190
Table 6.4	*HAH1* and *annexin* VI RT-PCR conditions	193
Table 6.5	*HAH1* and *annexin* VI cycle sequencing PCR conditions	194

Table 6.6	Clinical details of 5q- syndrome patients included in the study	196
Figure 6.2	Northern blot analysis of I.M.A.G.E. cDNA clone 416547	197
Figure 6.3	Physical map of the *SPARC, HAH1,* and *annexin* VI genes	199
Figure 6.4	$CD34^{+}$ expression analysis of the *annexin* VI gene	200
Figure 6.5	Sequence analysis of exon 10 of the *SPARC* gene	201
Figure 6.6	BestFit analysis of patient 1 and the published sequence from fragment 2 of the *annexin* VI gene	203

Chapter 7

Table 7.1	Clinical details of 5q- syndrome and AML patients included in the study	223
Table 7.2	Exon primer conditions for the *GSHPx-3* and ENSG00000145872 genes	224
Table 7.3	Exon primer conditions for the *PDGFRß* gene	225
Table 7.4	Exon primer conditions for 15/23 exons of the *MEGF1* gene	226
Table 7.5	Exon primer conditions for novel gene ENSG00000086589	227
Table 7.6	5' RACE PCR conditions for candidate gene *MEGF1*	233
Table 7.7	Genomic PCR primer conditions for candidate gene *MEGF1*	234
Figure 7.1	Agarose gel analysis of the 358bp product of *GSHPx-3* exon 1	238
Figure 7.2	Agarose gel analysis of the 449bp PCR product of *PDGFRß* exon 23b	239
Figure 7.3	Creation of a mixture of heteroduplexes and homoduplexes through hybridisation	240
Figure 7.4a	Graphical representation of the melting curve of *MEGF1* exon 19	241
Figure 7.4b	Graphical representation of the temperatures required to partially denature *MEGF1* exon 19	241
Table 7.8	Temperature predictions for DHPLC analysis on the WAVE™ system, for candidate gene *MEGF1*	243
Table 7.9	Temperature predictions for DHPLC analysis on the WAVE™ system, for candidate genes *PDGFRß* and ENSG00000086589	244

Table 7.10	Temperature predictions for DHPLC analysis on the WAVE™ system, candidate genes *GSHPx-3* and ENSG00000145872	245
Figure 7.5a	DHPLC chromatogram of exon 1 of the ENSG00000145872 novel gene	246
Figure 7.5b	DHPLC chromatograms of heterozygotes	247
Table 7.11	Frequency of polymorphisms identified from the coding exons and flanking intronic sequence of candidate gene *MEGF1*, by DHPLC, in patients with the 5q- syndrome/ AML and normal controls	249
Table 7.12	Frequency of polymorphisms identified from the coding exons and flanking intronic sequence of candidate genes *GSHPx-3*, *PDGFRß*, and ENSG00000086589 by DHPLC, in patients with the 5q- syndrome/AML and normal controls	250
Figure 7.6	Sequence analysis of exon 14 of the *MEGF1* gene	251
Figure 7.7a	DHPLC analysis of exon 23e of the *PDGFRß* gene	253
Figure 7.7b	Sequence analysis of exon 23e of the *PDGFRß* gene	254
Figure 7.8	BestFit analysis of patient 13 and the published sequence from exon 8 of the ENSG00000086589 novel gene	255
Figure 7.9	BlastX analysis of 5' RACE sequence and the *Homo sapiens* chromosome 5 working draft sequence, contig AC011374	257
Chapter 8		
Figure 8.1	Transcription map of the critical region of the 5q- syndrome	265

Chapter 1

Introduction

1.1 The Myelodysplastic syndromes

1.2 Molecular pathogenesis of MDS and leukaemia

1.2.1 Cytogenetic studies

1.2.1 Molecular studies

1.3 The 5q deletion in MDS and leukaemia

1.4 The 5q- syndrome

1.4.1 Cytogenetic studies

1.4.1 The 5q- syndrome critical region

1.5 The Human Genome Project

1.6 Human Chromosome 5

1.7 Expressed Sequence Tags

1.8 I.M.A.G.E. cDNA clones

1.9 Aims of the study

1.1 The Myelodysplastic syndromes

The myelodysplastic syndromes (MDS) are a heterogeneous family of haematologic disorders characterised by ineffective haematopoiesis (Gordon, 1999). These syndromes usually present in the elderly but can be seen in younger patients, including children, and are increasingly being recognised as a complication of chemotherapy used in the treatment of a variety of human malignancies (Passmore *et al.*, 1995).

MDS is classified into primary MDS (no known cause) and secondary MDS (strong association with a leukaemogenic agent). The latter usually have multiple chromosome abnormalities in the bone marrow and evolve to acute myeloid leukaemia (AML). A central criterion for the classification is the percentage of blasts, or primitive leukaemic cells, in the bone marrow. The French-American-British (FAB) classification system has developed five categories of disease: (1) refractory anaemia (RA), (2) refractory anaemia with ringed sideroblasts (RARS), (3) refractory anaemia with excess blasts (RAEB), (4) refractory anaemia with excess blasts in transformation (RAEB-t), and (5) chronic myelomonocytic leukaemia (CMML).

For the majority of patients with MDS, no curative option exists. Patients who are young enough and have an available matched sibling or matched unrelated donor may undergo an allogeneic bone marrow transplant (BMT) with a potential cure rate of 30% to 50% (Gordon, 1999). The major issue regarding this approach is the relatively high morbidity, or that no overall benefit will occur (relapse). The best results in terms of relapse-free survival appear to be in the subset of patients with early or low-grade MDS, characterised by RA or RARS. These patients that lack a donor for BMT are considered for induction chemotherapy. However, the

majority of elderly patients with MDS are not optimal candidates for such an approach. As a result, supportive care has a major role for patients with MDS and depending on the FAB presentation, may be the preferred approach. Erythropoietin, a growth factor, is probably the most commonly used supportive care after transfusion. The use of colony-stimulating growth factors to support leukopenia is currently under investigation.

1.2 Molecular pathogenesis of MDS and leukaemia

MDS is a clonal disorder that affects an early haematopoietic progenitor, giving rise to clonally derived neutrophils, erythrocytes and platelets. Evidence that MDS is a clonal disorder includes the presence of clonal cytogenetic abnormalities and analysis using X-inactivation-based clonality assays. These findings support the concept of somatic mutations giving rise to clonal proliferation of myeloid lineage cells. However, these studies provide no information about the identity of the genes involved in the development of myelodysplasia. The frequent loss of genetic material has led to the hypothesis that, by analogy with other malignancies characterised by genetic loss (e.g. retinoblastoma, Wilms' tumour, and colon cancer), MDS is caused by loss of function of a gene with tumour suppressor activity (Legare and Gilliland, 1995).

1.2.1 Cytogenetic studies

The reported frequency of karyotypic abnormalities in primary MDS varies between one third and one half of all successfully karyotyped cases (Second International Workshop on Chromosomes in Leukaemia, 1980). Approximately fifty percent of these abnormalities are due to the loss of genetic material as a result of whole or partial chromosome loss, specifically chromosomes 5, 7, 11, 12,

13, and 20 (Mufti, 1992). Monosomies and chromosome deletions have been reported in all five MDS subtypes, although some of these karyotypic abnormalities occur more frequently in specific subsets of the disease. For example, the 5q deletion is the most commonly reported abnormality in RA, and has been found in up to 70% of patients (Heim and Mitelman, 1986). This is in contrast with monosomy 7 that is found in only 5% of patients with RA, but 30% of patients with RAEB and RAEB-t, and 20% with CMML. The 11q deletion is frequently observed in RARS and is reported in up to 20% of patients. The 12p deletion is rarely reported in RA and RAEB, but is a frequent karyotypic abnormality in CMML (Mufti, 1992).

Cytogenetic studies have frequently identified non-random interstitial deletions and monosomy of chromosomes 5, 7, and 17 in MDS and AML suggesting a multistep pathway that culminates in an aggressive clinical course (Castro *et al.*, 2000). Thus, the unmasking of an oncogene(s) or inactivation of a tumour suppressor gene(s) on these deleted chromosomes may have an important role in the evolution of MDS to AML when they are mutated. In AML and MDS, mutations of the *p53* gene are infrequent, less than 10%. However, in a subset of patients with "17p- syndrome", the *p53* gene is frequently mutated (Lai *et al.*, 1995). A recent report on this distinct clinical entity demonstrated loss of the *p53* gene by 17p deletion in 14/16 cases analysed by FISH.

Deletion of the long arm of chromosome 20 (20q-) represents the most common chromosomal abnormalities associated with the myeloproliferative disorders (MPDs), but is also found in MDS and AML (Bench *et al.*, 2000). Bench *et al.*, have recently defined a MPD CDR of 2.7Mb, and a MDS/AML CDR of 2.6Mb. The authors localised twenty ESTs to the new MDS critical region of which five were

expressed in both human bone marrow and purified $CD34^+$ cells. These genes represent candidates for the 20q- MDS/AML gene.

Partial deletion of the long arm of chromosome 7, or monosomy 7, is a common abnormality in the bone marrow cells of patients with MDS or acute non-lymphocytic leukaemia (ANLL) (Kere *et al.*, 1987). Molecular studies using DNA markers that map to the terminal portion of 7q have shown the deletion to be interstitial, although terminal deletions have been reported (Fourth International Workshop on Chromosomes in Leukaemia, 1984). As for chromosome 20, studies have shown there to be more than one critical region of gene loss of the 7q-chromosome. The region 7q22-q34 may contain as many as four distinct minimal regions of deletion that are thought to contain one or more myeloid tumour suppressor genes (Todd *et al.*, 2001).

1.2.2 Molecular studies

The mechanism causing chromosomal deletions in MDS and leukaemia is unknown. Huebner *et al.*, (1989) suggested that specific genome rearrangements/deletions may be characteristic of and necessary for many differentiating lineages. That is, terminally differentiated cells within a particular lineage, rearrange specific chromosome regions in order to switch on or off the genes required for differentiation.

The consistent loss of genetic material in MDS and leukaemia is suggestive of a recessive mechanism of pathogenesis and it is probable that the deleted chromosome/bands harbour as yet unidentified tumour suppressor genes.

The importance of recessive mechanisms of tumourigenesis was first highlighted by Knudson in 1971. According to Knudson's 'two-hit' model, dominantly inherited predisposition to cancer entails a germline mutation, while tumourigenesis requires a second, somatic mutation. Non-hereditary cancer of the same type requires the same two hits, but both are somatic. The original tumour used in this model, retinoblastoma, involves mutation or loss of both copies of the *RB1* tumour suppressor gene in both hereditary and non-hereditary forms (Knudson, 1971).

Current molecular studies have identified at least seventeen tumour suppressor genes involved in the pathogenesis of cancer, including; *p53*, adenomatous polyposis coli (*APC*) gene, and the neurofibromatosis type 1 (*NF1*) gene. A review of the existing data about the various tumour suppressor genes and their role in disease leads to the conclusions that; the mutation of a single tumour suppressor gene can predispose to tumours in multiple tissues, and the products of tumour suppressor genes function at many levels within the cell (Skuse and Ludlow, 1995). For instance, they may be transcription factors, e.g. *RB1*, they may play a part in cytoplasmic signal transduction, e.g. *NF1*, or they may be cell-surface adhesion molecules, e.g. the deleted in colorectal cancer (*DCC*) gene. In every case, the tumour suppressor protein has a role in the normal regulation of cellular proliferation or differentiation. Tumours arise when that role is disturbed and cell growth becomes unrestrained.

There are many recent reports suggesting that known tumour suppressor genes may be involved in the pathogenesis of MDS and leukaemia. These include the *p53*, *p15* and *p16* genes. A study by Kaneko *et al.*, (1995) showed 7/57 patients with MDS demonstrated *p53* mutations within exons 5 to 8. Four of these patients progressed to acute leukaemia within seven months of diagnosis, and the

remaining three died within seven months without leukaemic transformation. These findings suggest that mutations of the *p53* gene can be implicated in leukaemic transformation and a poor prognosis in MDS (Kaneko *et al.*, 1995). Moreover, Adamson *et al.*, (1995) showed that 4/26 MDS patients had deletion of the *p53* gene in exons 5 to 8. Each case with a mutation was of an advanced MDS subtype, suggesting that *p53* mutation in these diseases is a terminal genetic event in the process of leukaemogenesis.

In the last five years, two cyclin-dependent kinase inhibitors, known as *p16*(INK4A/MTS1) and *p15*(INK4B/MTS2), which map to 9q21, have been found deleted in a wide range of tumours and mutated in a small number of leukaemic patients (Sill *et al.*, 1996). Moreover, recently it has been shown that the *p15* gene is methylated in MDS. Hypermethylation and homozygous deletions of tumour suppressor genes establish a new paradigm of inactivation by lack of expression, in contrast to the previously identified tumour suppressors which are predominantly inactivated by point mutations followed by loss of the wild-type allele. To investigate the time sequence of occurrence of *p15* gene methylation in MDS and its correlation with leukaemic transformation and survival of patients, Tien *et al.*, (2001) analysed the methylation status of the *p15* promoter region in fifty patients and was serially studied in twenty-two of them. 17/50 (34%) patients showed *p15* gene methylation, first demonstrated at diagnosis or during follow-up. When FAB subtypes at the time of study were used in the analysis, the incidence of *p15* methylation in each risk group remained stable throughout the course: 0% for RA and RARS, and from 23% to 30% for RAEB, RAEB-t and CMML. The incidence of *p15* methylation rose to 60% at initial study and finally, to 75% in cases of AML evolved from MDS. Most patients (69%) with *p15* methylation showed disease progression to AML. Thus, *p15* methylation can be detected early at the diagnosis of MDS or acquired during disease progression. It

may play an important role in the pathogenesis of some high-risk MDS and is related to leukaemic transformation of MDS (Tien *et al.*, 2001).

The isolation and characterisation of these putative tumour suppressor genes will lead to an understanding of molecular mechanisms underlying normal haematopoiesis and leukaemic transformation. Several strategies are currently in use to clone these genes, including defining a critical region of gene loss using molecular studies and FISH (Fluorescent *in situ* Hybridisation) analysis; constructing a YAC (Yeast Artificial Chromosome) contig giving genomic coverage of the whole region; isolating novel coding sequences from these YACs from appropriate cDNA libraries; screening expressed sequence tag (EST) databases; and more recently, using the draft and/or annotated sequence available from the Human Genome Project (HGP). The tumour suppressor gene is ultimately identified as a result of its frequent inactivation by rearrangement or point mutation. To date, however, no tumour suppressor genes have been definitively identified from the study of chromosome deletions in MDS or AML.

1.3 The 5q deletion in MDS and leukaemia

The 5q deletion is the most frequently reported deletion in MDS and is observed in 10-15% of patients (Heim and Mitelman, 1986). The deletion is interstitial, the breakpoints are variable, and all thirteen bands between 5q11 and 5q35 have been cited as breakpoints. The breakpoints most frequently reported are 5q12-14 (proximal) and 5q31q33 (distal). There also appears to be breakpoint uniformity with respect to the 5q deletion. Johansson *et al.*, (1993) and Pedersen and Jensen, (1991) found that the del(5)(q13q33) is the most commonly reported 5q deletion in MDS. The proposed location of the critical region of the 5q- chromosome differs markedly between cytogenetic studies suggesting the presence of more than one

critical region. This has been confirmed by molecular studies that have defined more than one critical region of the 5q- chromosome. Boultwood and Fidler, (1995 review) describe four distinct CDRs between chromosome 5q31-q33 reported in the literature. However, the approximate 5Mb CDR reported by Boultwood *et al.*, (1994a) was defined by patients with the 5q- syndrome, while Willman *et al.*, (1993); Le Beau *et al.*, (1993), and Nagarajan *et al.*, (1994) defined their 5q deletion CDRs with MDS, therapy related MDS, and AML patients. Willman *et al.*, defined a more centromeric critical region mapping between interleukin 5 (*IL5*) and the granulocyte/macrophage colony stimulating factor (*CSF2*) centering around the interferon regulatory factor 1 (*IRF1*) gene in a group of patients with MDS and AML. Le Beau *et al.*, have defined the critical region of the 5q deletion as the approximately 2.8Mb region between the interleukin 9 (*IL9*) gene and genetic marker D5S166 in a range of malignant myeloid disorders. Nagarajan *et al.*, defined their critical region between that of Willman and Le Beau, encompassing the early growth response (*EGR1*) gene. This suggests the disease genes causing the 5q- syndrome, therapy related MDS and AML may be distinct, and that chromosome 5q31-q33 harbours more than one tumour suppressor gene.

A number of haematopoietic growth factors and receptors have been localised to 5q an it has been speculated that one or more of them may be critical to the pathogenesis of these myeloid disorders. The heamatological role of these genes has led to the proposal that loss of one or more of these genes may be critical to pathogenesis in those myeloid disorders with a 5q deletion. They include the interleukins 3, 4, 5, and 9, *CSF2*, receptor for the macrophage colony stimulating factor (*CSF1R*), *EGR1*, and *IRF1* gene, see Figure 1.1. The molecular analysis of candidate and newly identified genes mapping within the respective critical regions should reveal the nature of the genetic abnormality associated with the myelodysplastic syndromes in which these deletions are found.

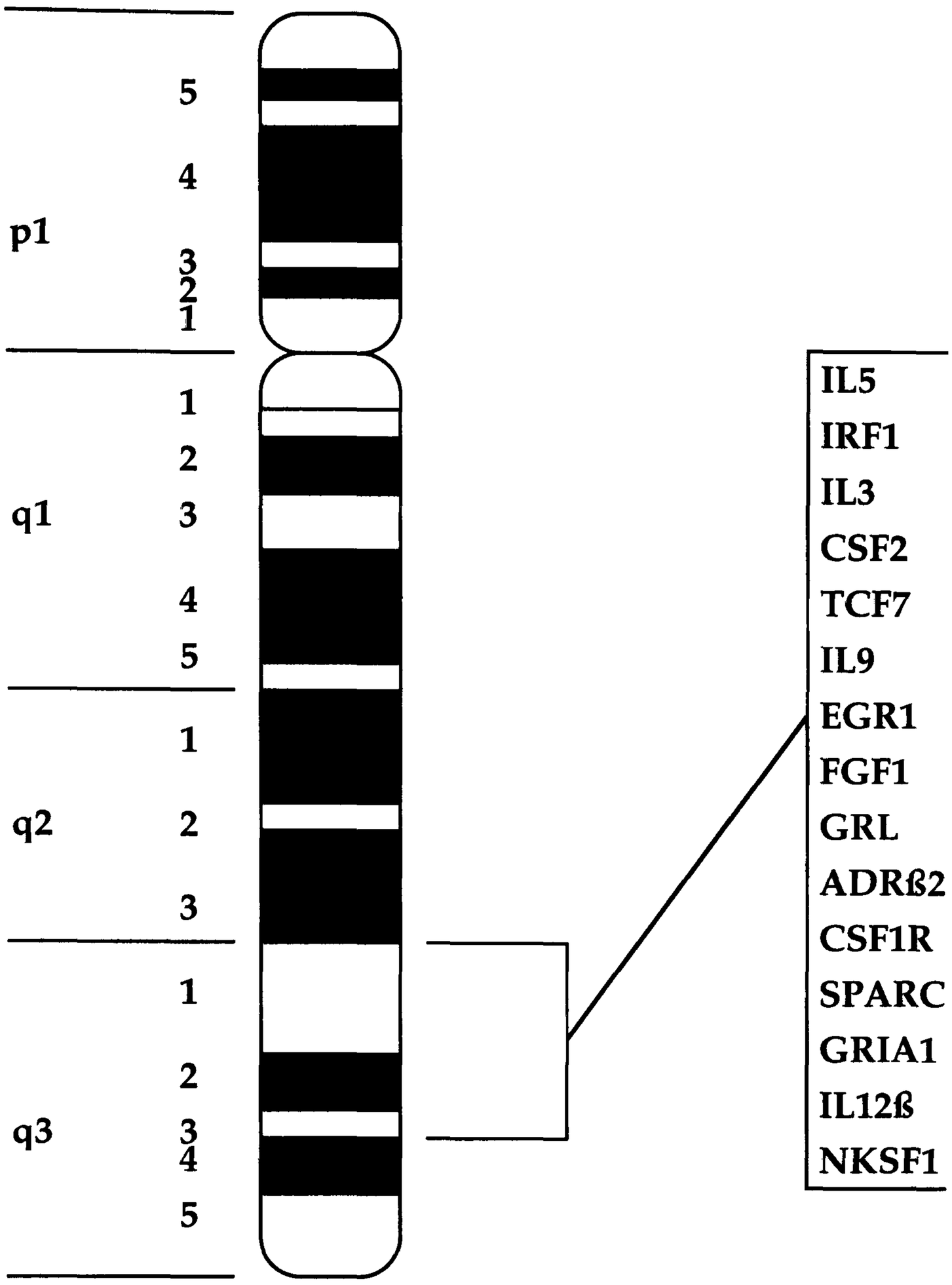


Figure 1.1

Ideogram of chromosome 5 showing some of the genes localised to 5q31-q33.

Early studies focused on the *EGR1*, *CSF1R*, and putative tumour suppressor gene, *IRF1*, as candidate genes for MDS and AML in association with a 5q deletion. A study by Willman *et al.*, (1993) showed *IRF1* to be consistently deleted at one or both alleles in thirteen cases of MDS/AML with aberrations of 5q31. However, a study by Boultwood *et al.*, (1993) showed the *IRF1* gene to be retained in patients with the 5q- syndrome. The *EGR1* gene was also shown to be retained on the 5q- chromosome in patients with the 5q- syndrome, thus mapping it outside the

critical region of gene loss (Boultwood *et al.*, 1994a). The *CSF1R* gene is a transmembrane glycoprotein with tyrosine kinase activity (Sherr *et al.*, 1985). The loss of one *CSF1R* allele from the 5q- chromosome together with the loss of the second allele on the apparently normal homologous chromosome 5 in a subpopulation of cells in some patients with MDS was demonstrated by Boultwood *et al.*, (1994a).

1.4 The 5q- syndrome

The 5q- syndrome was first described by Van den Berghe *et al.*, in 1974 when three patients with long-standing idiopathic refractory anaemia and an interstitial deletion of the long arm of chromosome 5 were discovered. Two additional patients were reported in 1975 by Sokal *et al.*, and a new haematologic syndrome was established, the 5q- syndrome. Patients with the 5q- syndrome are typically elderly (mean age at presentation is 66 years), predominantly female, present with macrocytic anaemia, modest leukopenia, normal or high platelet count, a hypercellular bone marrow, hypolobulated megakaryocytes, del (5q) as the sole karyotypic abnormality, and a low risk of transformation to acute leukaemia. The management of patients with the 5q- syndrome is supportive. The major manifestation of the syndrome relates to the refractory anaemia and the need for RBC transfusions. Patients with severe anaemia who receive repeated red cell transfusions are at risk for the infectious complications of blood transfusions, as well as iron overload.

1.4.1 Cytogenetic studies

The 5q deletion occurs as the sole karyotypic abnormality in the 5q- syndrome, as well as together with other karyotypic abnormalities in the other MDS subtypes

RAEB, RARS, secondary MDS, and *de novo* AML. In contrast to t-MDS and t-AML with del(5q), the 5q- syndrome usually has a benign clinical course, with a low risk of transformation to acute leukaemia. However, the del(5q) in the 5q- syndrome is cytogenetically indistinguishable from the deleted chromosome 5 of other myeloid disorders (Jaju *et al.*, 1998).

The mechanism causing the 5q- syndrome is unknown. The frequent loss of genetic material from the long arm of chromosome 5 in association with the 5q- syndrome is suggestive of a recessive mechanism of tumourigenesis, and it is probable that the deleted chromosome bands harbour a tumour suppressor gene (Boultwood *et al.*, 1994a).

1.4.2 The 5q- syndrome critical region

The 5q deletion typically encompasses most of the long arm of chromosome 5, del(5)(5q13q33), and many known genes are lost as a result (Boultwood *et al.*, 1994a). However, uncharacteristically small 5q deletions in association with MDS have been reported (Van den Berghe *et al.*, 1985). Prior to this study, Boultwood *et al.*, had identified two patients with the 5q- syndrome and small 5q deletions, del(5)(5q31q33). Loss of heterozygosity (LOH) analysis was carried out on a large number of genes localised to this region, with the aim of identifying the gene(s) involved in the pathogenesis of the 5q- syndrome. The CDR of 5.6Mb was delineated between the gene for fibroblast growth factor acidic (*FGF1*) and the subunit of interleukin 12 (*IL12ß*) in these two patients with the 5q- syndrome and small deletions. The common region of loss in these two 5q- syndrome patients include the glucocorticoid receptor (*GRL*), adrenergic receptor beta 2 (*ADRß2*), *CSF1R*, secreted protein, acidic, and rich in cysteine (*SPARC*), and glutamate receptor (*GLUH1*) genes. These genes were shown to be deleted from the 5q-

chromosome in both patients, and thus represent candidates for the 5q- syndrome gene. This 5q- syndrome critical region is telomeric to and distinct from the other critical regions on 5q associated with MDS and AML.

1.5 The Human Genome Project

The Human Genome Mapping Project initiated in 1990 is a thirteen year, $13 billion effort coordinated by the US Department of Energy (DOE) and the National Institute of Health (NIH). The project was originally planned to last fifteen years, but rapid technological advances have accelerated the expected completion date to 2003, as of January 31, 2001. The project goals are to:

- *identify* the 100,000 genes in human DNA (now believed to be 30,000).
- *determine* the sequences of the 3 billion bp that make up human DNA.
- *store* this information in databases.
- *develop* tools for data analysis.
- *address* the ethical, legal and social issues that may arise from the project.

There are five International Human Genome Project major sequencing sites: DOE national laboratories; Baylor College of Medicine Genome Centre; The Sanger Centre; The Washington University Genome Sequencing Centre; and The Whitehead Institute/MIT Centre for Genome Research. There are twenty-one other US Genome Research sites working on individual chromosomes and fifteen other International Genome Research Centres from many countries including Australia, China, France, Germany, Japan, and Korea. The five major sites (from the US and UK) have split the project between them: The Genome Centre at Washington University are sequencing chromosomes 1, 2, 3, 4, 5, 7, 8, 12, 13, 14, 16, 22, and X. The Sanger Centre are sequencing chromosomes 1, 6, 9, 10, 13, 20, 22, and X. The first major breakthrough was the recent completion of the 33.4Mb sequence of chromosome 22 (Dunham *et al.*, 1999). As of November 10 2001, over

671,627,342 bp (22%) of the 3 billion bp is finished, high-quality sequence. A further 499,693,897 bp (16%) is unfinished sequence.

1.6 Human Chromosome 5

Human Chromosome 5 spans 198cM and contains 194Mb of DNA and represents approximately 6% of the human genome. It is currently being sequenced by Washington University Genome Centre. On April 13 2000, researchers had decoded in draft form the genetic information on chromosome 5. As of October 31 2001, 22.4Mb (11.5%) of the 194Mb is finished, high-quality sequence.

1.7 Expressed Sequence Tags

Until 1991, the field of human genome research had predominantly concerned itself with large-scale mapping and technology development, for example, positional cloning and the advent of automated sequencing. Many exciting insights have arose from positional cloning: the involvement of trinucleotide repeats in the pathophysiology of Fragile X syndrome, Huntington's disease and other neurological disorders is a striking example (La Spada *et al.*, 1994). However, up until 1995, only approximately forty genes had been cloned by virtue of their location in the genome, compared with approximately 36,000 sequences in the primate division of GenBank that are the result of functional cloning experiments. Thus, in 1991, the introduction of expressed sequence tags (ESTs), partial 5' and 3' cDNA sequences (approximately 300-400bp) representing expressed human genes, was considered the route to completing the 3 billion nucleotide sequence of the human genome. However, it was not until 1994 with the advent of the Washington University EST Project, funded by Merck and Co.

and the National Cancer Institute, that ESTs gave the public data collections a boost.

The National Centre for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov) was established on November 4 1988. It is located at the National Library of Medicine, on the campus of the NIH. It is responsible for building, maintaining, and distributing biomedical databases including GenBank, and the NIH genetic sequence database that collects all known DNA sequences from scientists worldwide. GenBank (http://www.ncbi.nlm.nih.gov/Genbank/index.html) is the NIH genetic sequence database; an annotated collection of all publicly available DNA sequences. There are approximately 14,397,000,000 bases in 13,602,000 sequence records as of October 2001. One division of GenBank is the EST database db (EST) (http://www.ncbi.nlm.nih.giv/dbEST/index.html). On November 2 2001, the number of public entries in db(EST) was 9,407,866, of which 3,876,441 were human. db(EST) has had a profound effect on the positional candidate approach of gene discovery. As of December 5 1997, 91% (83/91) of positionally cloned genes mutated in human disease states were represented by exact matches with one or more ESTs in db(EST).

Several research groups have utilised the EST resource in assigning expressed genes to specific chromosomes and regions. ESTs that are regionally assigned can serve as sequence tagged sites (STSs) for physical mapping projects, and have the advantage of representing an expressed gene. Furthermore, when localised, ESTs can serve as candidate genes for disease loci in the region. As the human expression map becomes denser, the use of ESTs as candidate genes may supplement the traditional strategy of positional cloning (Pappas *et al.*, 1995). In 1993, Sargent *et al.*, identified the gene for glycerol kinase deficiency through the

use of EST localisation as well as a traditional positional cloning approach. Recently, ESTs have been localised to transcript maps on several chromosomes, including chromosome 5, with the aim of identifying candidate tumour suppressor genes for cancer and genetic disease. For example, Horrigan *et al.*, (1999) constructed a high-resolution map of a 6Mb interval of human chromosome 5, band 5q31, incorporating 175 STSs, of which 122 were non-redundant ESTs. The ESTs were assembled into overlapping transcription units and ordered with respect to polymorphic markers in the region, resulting in a comprehensive map. This map will facilitate gene discovery efforts for several disorders, including the genes deleted in acute myeloid leukaemias and myelodysplasia (Horrigan *et al.*, 1999). The distal short arm of chromosome 1 (1p) is rearranged in a variety of malignancies, and several genetic diseases also map to this region (Jensen *et al.*, 1997). Jensen *et al.*, constructed an integrated transcript map to precisely define the positions of genes and ESTs previously mapped to 1p35-p36. One hundred and forty-two ESTs were mapped by PCR against a radiation hybrid panel, with the aim of identifying candidate genes for genetic disease mapping to distal 1p.

1.8 I.M.A.G.E. cDNA clones

The Integrated Molecular Analysis of Genomes and their Expression (I.M.A.G.E.) consortium at http://www-bio.llnl.gov/bbrp/image/image.html was initiated in 1993 by four academic groups on a collaborative basis after informal discussions led to a common vision on how to achieve an important goal in the study of the human genome. The groups shared high-quality, arrayed cDNA libraries and placed sequence map and expression data on the clones in these arrays into the public domain. From this information, they re-arrayed the unique clones to form a "master array" which they hoped would ultimately contain a representative cDNA from each and every gene in the genome. There are five authorised

distributors of I.M.A.G.E. clones from either the USA: American Type Culture Collection (ATCC); Genome Systems Inc.; Research Genetics; or Europe: UK HGMP, Hinxton, England; and Resource Centre of the German Human Genome Project, Berlin Germany. The majority of ESTs deposited into db(EST) are from the I.M.A.G.E consortium and each one is represented by an I.M.A.G.E. cDNA clone. These clones are available to researchers free of any royalties via the world-wide-web (www).

1.9 Aims of the study

Following the delineation of the critical region of the 5q- syndrome to approximately 5Mb at 5q31-5q33, the primary aim of this study was to generate a transcript map of the CDR of gene loss, flanked by the *FGF1* and *IL12ß* genes. The development of a transcript map should facilitate the identification of candidate tumour suppressor genes on 5q.

As previously mentioned, the majority of genes in the human genome are represented by one or more ESTs. We decided to use db(EST) as the primary resource to identify novel coding sequences mapping to the critical region of gene loss. Molecular analysis including mutation studies was carried out on candidate genes, with the ultimate aim of identifying the 5q- syndrome gene.

Chapter 2

Methods

2.1 Preparation of Granulocyte and T-lymphocyte fractions from whole peripheral blood

2.1.1 Separation of granulocytes and mononuclear cells by density gradient centrifugation

2.1.2 Preparation of mononuclear cells

2.1.3 Separation of T-lymphocytes by rosetting with sheep red blood cells

2.2 Extraction of high molecular weight DNA from whole peripheral blood, and peripheral blood cell fractions.

2.2.1 Cell preparation for whole blood

2.3 Standard restriction enzyme digestion of genomic DNA and gel electrophoresis

2.4 Southern blotting

2.5 Preparation of probes for hybridisation

2.5.1 Genomic and cDNA probes

2.5.1.1 Transformation of competent cells

2.5.1.2 Plasmid DNA mini-preparation

2.5.1.3 Recovery of probe from plasmid

2.5.2 PCR generated probes

2.6 Probe labelling

2.7 Filter hybridisation

2.8 Filter washing

2.9 Removal of the probe from the filter

2.10 Autoradiography

2.11 Preparation of single-stranded templates for DNA sequencing

2.11.1 Cloning of the insert from the plasmid into M13 phage

2.11.1.1 Preparation of insert and vector DNA

2.11.1.2 Ligation of insert into the M13 vector

2.11.2 Preparation of competent *E.coli* for transformation

2.11.3 Transformation of competent cells with ligated M13

2.11.4 Preparation of single-stranded templates

2.12 Automated fluorescent dye sequencing

2.12.1 Sequencing reactions using a Cy5 labelled primer

2.12.1.1 Annealing of the primer to a single-stranded template

2.12.1.2 Sequencing reactions

2.12.1.3 Annealing of the primer to a double-stranded template

2.12.1.4 Sequencing reactions

2.12.2 Sequencing reactions using a Cy5 dATP internal label

2.12.2.1 Annealing of the primer to a single-stranded template

2.12.2.2 Sequencing reactions

2.12.2.3 Annealing of the primer to a double-stranded template

2.12.2.4 Sequencing reactions

2.12.3 Preparation of the sequencing gel plates

2.12.4 Preparation of the gel

2.13 RACE PCR

2.14 Subcloning of RACE PCR products for direct sequencing

2.14.1 A-Tailing of RACE PCR product

2.14.2 Ligation of RACE PCR products into the vector

2.14.3 Transformation of competent cells

2.15 Localisation of known genes and novel cDNAs to the YAC contig

2.16 Reverse transcriptase PCR (RT-PCR) analysis

2.16.1 Purification and quantification of RT-PCR products.

2.17 $CD34^+$ expression by RT-PCR

2.18 Cycle sequencing on the ALF*express* automated sequencer

2.18.1 Preparation of dNTP/Cy5 ddNTP mixes

2.18.2 Sequencing reactions

2.18.3 Precipitation of sequencing reactions

2.18.4 Preparation of the polyacrylamide gel and loading of samples

2.18.5 Processing the sequence data on the ALF*express* automated sequencer

2.19 Cycle sequencing on the ABI PRISM 3100 Genetic Analyser

2.19.1 Preparation of sequencing reactions

2.19.2 Sodium acetate precipitation of sequencing reactions

2.19.3 Processing the sequence data on the ABI PRISM 3100 Genetic Analyser

The preparation of all reagents used in this Methods section is described in the Appendix.

2.1 Preparation of Granulocyte and T-lymphocyte fractions from whole peripheral blood.

The method described is for the separation of granulocytes and T-lymphocytes from 20mls of peripheral blood. Sample volumes of 40mls were routinely separated in 2 x 20mls.

2.1.1 Separation of granulocytes and mononuclear cells by density gradient centrifugation.

1. 20mls of peripheral blood was collected into Tri-Sodium EDTA tubes.
2. The blood was gently layered onto 20mls of Histopaque-1077 (Sigma Aldrich, Cheshire, UK) in a 50ml polypropylene conical tube and centrifuged at 1600rpm (400g, Sorvall RT6000B benchtop centrifuge) for 30 minutes at room temperature.
3. The interface (the mononuclear cell layer) was transferred to a sterile 50ml conical tube, using a Pasteur pipette, and processed as described (**2.1.2**).
4. The upper (serum) and lower (Histopaque) layers were carefully aspirated using a Pasteur pipette and discarded to leave the red blood cell/granulocyte layer.
5. Phosphate buffered saline (PBS) was added to the red blood cell/granulocyte layer to give a total volume of 50mls and the tube mixed by gentle inversion.
6. The tube was centrifuged at 1600rpm at room temperature for 10 minutes.
7. The supernatant was removed using a Pasteur pipette and discarded.
8. Steps 5, 6 and 7 were repeated.
8. To lyse the red cells, the packed red blood cell/granulocyte layer was distributed into conical tubes containing freshly prepared red cell lysis buffer (approximately 1ml of red blood cells per 50ml of red cell lysis buffer), and left at room temperature for 15 minutes with occasional mixing.

8. The tubes were centrifuged at 1600rpm for 10 minutes at room temperature and the supernatant poured off.
8. Each granulocyte pellet was resuspended in approximately 1ml of PBS and the pellets pooled into 2 conical tubes.
8. The tubes were filled to 50mls with PBS and centrifuged at 1600rpm for 10 minutes at room temperature. The supernatant was poured off.
8. Step 12 was repeated.
8. The pellets were pooled into one conical tube and finally resuspended in PBS to a total volume of 10mls.
8. If the cell fraction was for the preparation of high molecular weight DNA, DNA was either extracted immediately, see section **2.3** from step 8, or the cell suspension was frozen at -20°C for DNA extraction at a later date.

2.1.2 Preparation of mononuclear cells

1. PBS was added to the mononuclear cell fraction to a total volume of 50mls, it was mixed by gentle inversion and centrifuged at 1600rpm for 10 minutes at room temperature.
2. The supernatant was poured off and the cell pellet resuspended in 50mls of PBS and centrifuged at 1600rpm for 10 minutes at room temperature.
3. The cell pellet was resuspended in 10mls of PBS.
4. This cell fraction was then either processed further to obtain T-lymphocytes, see section **2.1.3**, or used for the preparation of high molecular weight DNA, see section **2.2**

2.1.3 Separation of T-lymphocytes by rosetting with sheep red blood cells

This is based on the erythrocyte rosetting method of Kaplan and Clark (1974).

1. The mononuclear cell fraction from section **2.1.2** step 3 was diluted with PBS to obtain a concentration of 2-6 x 10^6/ml white blood cells.
2. 1-2 volumes of neuramidase-treated sheep red blood cells (TCS Biologicals, Buckingham, UK), 0.5-1 volumes of foetal calf serum (FCS) and 100-300µl of a fresh 1:30 dilution of a 1% stock solution of polybrene were added to the cell suspension.
3. The suspension was centrifuged at 750rpm for 5 minutes at 4°C and then incubated at 4°C for a minimum of 5 hours or maximum overnight.
4. The supernatant was removed from the packed cell pellet (sheep red blood cells and rosetted T-lymphocytes) and 1 volume of PBS added.
5. The cells were gently resuspended by rotating the meniscus through the cell pellet.
6. The cell suspension was gently layered onto an equal volume of Histopaque and centrifuged at 1600 rpm for 30 minutes at room temperature.
7. The upper layer, interface (non T-cell) and Histopaque layer were aspirated and discarded, leaving the sheep red blood cell/T-lymphocyte layer.
8. PBS was added to 50mls, the tube mixed by gentle inversion and centrifuged at 1600 rpm for 10 minutes at room temperature.
9. The supernatant was removed using a Pasteur pipette and discarded.
10. Steps 8 and 9 were repeated.
11. To lyse the sheep red blood cells, the cell pellet was distributed into conical tubes containing freshly prepared red cell lysis buffer (approximately 1ml of cells per 50ml of lysis buffer) and incubated at room temperature for 15 minutes with occasional mixing.
12. The tubes were centrifuged at 1600 rpm for 10 minutes and the supernatant poured off.
13. The T-cell pellets were treated exactly as the granulocyte pellets from step 11 of section **2.1.1**.

2.2 Extraction of high molecular weight DNA from whole peripheral blood, and peripheral blood cell fractions

Extraction was carried out using the Nucleon® BACC2 Genomic DNA Extraction Kit (Nucleon® Biosciences, Scotlab, Lanarkshire, UK). Steps 1-5 were omitted for DNA extraction from blood cell fractions.

2.2.1 Cell preparation from whole blood

1. Whole peripheral blood was collected in sodium EDTA tubes.
2. Samples were centrifuged at 2400rpm (1300g) for 10 minutes at room temperature and the plasma pipetted off and discarded, taking care not to disturb the buffy coat. The sample could be frozen at -20°C at this stage for extraction at a later date.
3. The sample was transferred to a 50ml polypropylene centrifuge tube and Reagent A (Nucleon® Biosciences) added to 40mls.
4. The sample was vortexed for 4 minutes and centrifuged at 2400rpm for 4 minutes at room temperature.
5. The supernatant was discarded without disturbing the pellet.
6. 2mls of Reagent B (Nucleon® Biosciences) was added to the cell pellet and vortexed briefly to resuspend the pellet.
7. The cell suspension was transferred to a 5ml screw-capped polypropylene centrifuge tube (maximum internal diameter 12mm).
8. 500µl of sodium perchlorate (Nucleon® Biosciences) was added and the tube inverted 10 times by hand.
9. 2mls of chloroform was added and the tube inverted 10 times by hand to emulsify the phases.
10. 300µl of Nucleon® resin was added, and without re-mixing the phases, centrifuged at 2400rpm for 3 minutes at room temperature.
11. The upper phase was transferred to a fresh 15ml polypropylene centrifuge tube without disturbing the Nucleon® resin layer (brown in colour).

12. Two volumes of cold absolute ethanol were added to the upper phase and the tube inverted several times until the DNA had precipitated. The sample can be stored at -20°C for >1 hour to aid precipitation of the DNA if necessary.
13. The sample was centrifuged at 3000rpm for 5 minutes to pellet the DNA and the supernatant discarded.
14. 2mls of cold 70% ethanol was added and the tube inverted several times. The sample was re-centrifuged as before, and the supernatant discarded. This step can be repeated if necessary.
15. The pellet was air dryed for 15 minutes and resuspended in an appropriate volume of sterile water. The DNA was left to dissolve overnight at 4°C and stored at -20°C.

2.3 Standard restriction enzyme digestion of genomic DNA, and gel electrophoresis

1. Approximately 5µg of DNA was digested in a total volume of 30-50µl.
2. The following were added to a 1.5ml microcentrifuge tube;
   i. Sterile distilled water to give the desired total volume.
   ii. The appropriate volume of 10x concentrated enzyme buffer to give a final 1x concentration. Specific buffers supplied by the enzyme manufacturers were used.
   iii. 100mM Spermidine (Sigma Aldrich).
   iv. 5µg of DNA.
   v. 40 units of restriction enzyme.
3. The contents of the tube were mixed gently and centrifuged briefly at low speed to bring the contents to the bottom of the tube.
4. The mixture was incubated at the optimal temperature for the restriction enzyme for a minimum of 4 hours.
5. In some cases it was necessary to add more enzyme. A total volume of 7-14µl of sterile distilled water, 10x enzyme buffer to a final concentration of 1x,

100mM Spermidine, and 40-80 units of enzyme was added to the digested DNA and steps 3 and 4 repeated.

6. Digested DNA was either electrophoresed immediately or stored at -20°C until required.
7. 1% agarose gels were prepared in a total volume of 300ml (the appropriate volume for the electrophoresis equipment used).
8. The agarose (Type I: Low EEO, Sigma Aldrich) was dissolved in 300ml of 1x TBE buffer and microwaved on high power for approximately 5 minutes or until the agarose was dissolved.
9. The molten agarose was poured into a sealed gel tray, with a 20-toothed comb positioned 1.0 cm from one end, and allowed to set.
10. The comb and tape were removed and the gel tray positioned in the electrophoresis tank with 2 litres of 1x TBE buffer.
11. Loading dye was added to the digested DNA samples (the volume of dye added was 1/10 of the volume of the digested DNA sample) and the samples loaded into the wells of the gel.
12. The gel was run at 50 volts overnight.

2.4 Southern blotting

1. Following electrophoresis, the gel was stained in ethidium bromide (Sigma Aldrich) (10mg/ml in distilled water) for 15 minutes, and then destained in distilled water for 10 minutes.
2. The gel was viewed on a UV transilluminator and a photograph taken (Polaroid film type 667,Fahrenheit, Milton Keynes, UK).
3. The gel was soaked in denaturing solution for 30 minutes and then immersed in alkali transfer buffer for 10 minutes.
4. A capillary blot was set up using alkali transfer buffer;
    i. A glass tray was filled with alkali transfer buffer, to form a reservoir.

ii. A glass plate was placed over the tray and a length of Whatman 3mm paper, soaked in alkali transfer buffer, was placed over the glass plate so that both ends of the paper were immersed in the buffer reservoir.

iii. The gel was inverted, placed on the Whatman paper and the edges of the tray covered in clingfilm to prevent evaporation of the buffer reservoir.

iv. A piece of Hybond N^+ membrane (Amersham Pharmacia Biotech, Amersham, UK) was cut to the size of the gel, numbered, soaked in distilled water and placed number side up on the gel.

v. 3 sheets of Whatman 3mm paper, cut to size, were soaked in alkali transfer buffer and placed on top of the filter.

vi. A whole pack of paper towels was placed on top of the Whatman paper and finally a second glass tray was inverted on top of the paper towels to provide weight.

5. Transfer was allowed to proceed for 16-24 hours.

6. The filter was soaked in neutralisation solution for 10 minutes and baked at 80°C for 10 minutes.

7. DNA was fixed to the filter by UV cross-linking, it was placed DNA side down on a transilluminator (wavelength 302nm) for 1 minute 30 seconds. The UV transilluminator was calibrated by cutting a Southern blot of control DNA into 5 strips and exposing each strip DNA side down on the transilluminator for a different length of time, ranging from 30 seconds to 5 minutes. The strips were all hybridised together with a suitable labelled probe. Following autoradiography, the optimum UV exposure time was evident from the strength of the signals.

2.5 Preparation of probes for hybridisation

2.5.1 cDNA probes

cDNA probes were generated from I.M.A.G.E. cDNA clones for use in Southern and Northern blot analysis.

2.5.1.1 Transformation of competent cells

1. Competent cells (Library Efficiency TM HB101, Gibco Life Technologies, Paisley, UK) were thawed on ice.
2. 1µl of plasmid DNA (10ng/µl) was mixed with 50µl of competent cells and incubated on ice for 30 minutes.
3. The tube was heat shocked in a water bath at 42°C for 90 seconds and returned to ice for 2 minutes.
4. 950µl of sterile LB (Luria Bertani) medium was added and the mix incubated at 37°C for 1 hour.
5. An appropriate volume (100µl for 15mm x 100mm plates) was spread onto solid LB agar plates containing the appropriate antibiotic.
6. The plates were inverted and incubated at 37°C overnight.
7. A single isolated colony was picked from the plate using a sterile loop and used to inoculate 10mls of sterile LB containing the appropriate antibiotic.
8. The bacteria were cultured overnight at 37°C with shaking.
9. A permanent stock was prepared by adding 0.15ml of glycerol to 0.85ml of the overnight culture in a sterile microcentrifuge tube, and stored at -70°C

2.5.1.2 Plasmid DNA mini-preparation

Plasmid was extracted with the QIAprep® Spin Miniprep Kit (QIAGEN, Southampton, UK) using a microcentrifuge.

1. 1.5mls of the overnight 10ml culture containing the appropriate antibiotic was transferred to a 1.5ml microfuge tube. The remaining culture was centrifuged for 15 minutes at 4°C to obtain a bacterial pellet, and stored at -20°C.
2. 250µl of Buffer P1 (QIAGEN) was added to the bacterial cell pellet and vortexed to resuspend the cells.
3. 250µl of Buffer P2 (QIAGEN) was added, and the tube inverted 6 times to mix.
4. 350µl of Buffer N3 (QIAGEN) was added, and the tube inverted immediately but gently 6 times to mix.

5. The sample was centrifuged at 13,000rpm for 10 minutes at room temperature. During centrifugation, a QIAprep spin column was placed in a 2ml collection tube.
6. The supernatant from step 5 was pipetted into the QIAprep spin column and centrifuged at 13,000rpm for 1 minute at room temperature, and the flow-through discarded.
7. The QIAprep spin column was washed by adding 0.5mls of Buffer PB (QIAGEN), centrifuged at 13,000rpm for one minute at room temperature, and the flow-through discarded. This step is necessary to remove trace nuclease activity when using *endA*$^+$ strains such as the JM series. Host strains such as XL-1 Blue and DH5α™ do not require this additional wash step.
8. The QIAprep spin column was washed by adding 0.75mls of Buffer PE (QIAGEN), centrifuged at 13,000rpm for 1 minute at room temperature, and the flow-through discarded.
9. The QIAprep spin column was centrifuged as before for an additional 1 minute to remove residual wash buffer.
10. The QIAprep spin column was placed in a fresh 1.5ml microfuge tube. The DNA was eluted by adding 50µl of sterile water to the centre of the QIAprep spin column, left for 1 minute, centrifuged at 13,000rpm for 1 minute at room temperature, and stored at -20°C.

2.5.1.3 Recovery of the probe from the plasmid

The probe was recovered from the plasmid with the Wizard® PCR Preps DNA Purification System (Promega, Madison, WI).

1. 15µl of plasmid DNA was digested with the enzyme(s) required to excise the insert, (see section **2.4** for digestion of DNA).
2. A 1% Low Melting Temperature agarose gel was made using a comb which could accommodate the larger volumes of plasmid digests. The duration of electrophoresis was dependent on the sizes of fragments to be separated.

3. The digested plasmid DNA solution was mixed with loading dye and loaded into the gel. A DNA molecular weight marker was loaded on one side.
4. Following electrophoresis, the gel was viewed on a UV transilluminator.
5. The band which corresponded to the correct size of the plasmid insert was cut from the agarose (approximately 300µl (300mg)) using a scalpel blade.
6. The 300µl (300mg) agarose slice was transferred to a 1.5ml microcentrifuge tube and incubated at 70°C until the agarose had completely melted.
7. 1ml of resin was added to the melted agarose slice and mixed thoroughly for 20 seconds by inverting.
8. One Wizard® Minicolumn was prepared. The plunger from a 3ml Luer-Lok® syringe (Becton Dickinson & Co., Oxford, UK) was removed and set aside. The syringe barrel was attached to the Luer-Lok® extension of each Minicolumn.
9. The resin/DNA mix from step 7 was pipetted into the syringe barrel. The syringe plunger was inserted slowly and the slurry gently pushed into the Minicolumn.
10. The syringe was detached from the Minicolumn, and the plunger removed from the syringe. The syringe barrel was reattached to the Minicolumn. 2mls of 80% isopropanol was pipetted into the syringe to wash the column. The syringe plunger was inserted into the syringe, and the isopropanol was gently pushed through the Minicolumn.
11. The syringe was removed and the Minicolumn transferred to a 1.5ml microcentrifuge tube. The Minicolumn was centrifuged at 5500rpm for 2 minutes at room temperature to dry the resin.
12. The Minicolumn was transferred to a new microcentrifuge tube. 50µl of sterile water was applied to the Minicolumn and allowed to stand for 1 minute. The Minicolumn was centrifuged at 5500rpm for 20 seconds to elute the bound DNA fragment.
13. The Minicolumn was removed and discarded, and the purified DNA quantified on a 1% TAE agarose gel and stored at -20°C.

2.5.2 Oligonucleotide probes

The oligonucleotide probes were generated by PCR amplification using primer sets designed from the obtained sequence data for Southern and Northern blot analysis.

1. For each 100µl PCR reaction the following were added to a sterile 0.6ml thin-walled tube;

sterile distilled water	up to 100µl
10x reaction buffer	10µl
50mM $MgCl_2$	6µl
dNTP mix (1.25mM)	8µl
primer 1 (100pmol)	1µl
primer 2 (100pmol)	1µl
template DNA (≈200ng)	1µl
Taq polymerase (2.0-2.5 units)	1µl

The buffer, $MgCl_2$, and *Taq* polymerase were manufactured by Bioline, London, UK). The dNTP mix (Amersham Pharmacia Biotech) was prepared by mixing 12.5µl of each of the 100mM dNTPs with 950µl of sterile distilled water and storing in 1ml aliquots. The volume of 50mM $MgCl_2$ stated was a standard starting point, if necessary a $MgCl_2$ titration was performed to determine the optimal final concentration. The template DNA was total genomic DNA obtained from the peripheral blood of normal healthy individuals.

2. The tubes were mixed gently and centrifuged briefly to bring the contents to the bottom of the tube.
3. Approximately 60µl of mineral oil was added to each tube, to prevent any evaporation, and the tubes centrifuged briefly again.
4. The tubes were placed on the thermal cycler (Biometra Trio Thermoblock, Anachem, Luton, UK) and subjected to the appropriate thermal profile. Each

primer set had an optimal annealing temperature and the thermal profile selected for each primer set was essentially as shown below, but the optimal annealing temperature was substituted for the annealing temperature shown.

Initial denaturation	95°C for 5 minutes
35 cycles of;	94°C denaturation for 30 seconds
	60°C primer annealing for 30 seconds
	72°C primer extension for 1 minute
Final extension	72°C for 8 minutes
Hold temperature	4°C

5. The mineral oil was removed from the tubes and a small aliquot of the reaction (usually 5µl) was run on a mini agarose gel. The concentration of the agarose was dependent on the size of the PCR product.
6. The gel was stained in ethidium bromide and visualised on a UV transilluminator. If a single band of the expected size was observed, then the remainder of the PCR product was purified. Purification was achieved by the use of the Wizard® PCR Preps DNA Purification System.
7. The purified product was quantified and stored at -20°C for use as a probe.

2.6 Probe labelling

Probes were labelled to high specific activity (approximately $2x10^9$ dpm/µg) with High Prime (Roche Products Ltd., East Sussex, UK) and radioactive dCTP (Amersham Pharmacia Biotech) using random oligonucleotide primers.

1. 25-40ng of probe, in distilled water to a final volume of 13µl, was denatured by boiling for 7 minutes and then quenched on ice for 3 minutes. For simultaneous hybridisation with two probes, the two probes were mixed in the desired ratio in a total volume of 13µl and the procedure followed as for a single probe.

2. The following were added to the denatured probe;
    i. 4µl of High Prime solution.
    ii. 3µl of [α^{32}P]dCTP, 3000Ci/mMol, aqueous solution.
3. The mixture was incubated at 37°C for 10 minutes.
4. Non-incorporated deoxyribonucleoside triphosphate was removed by chromatography through a Sephadex G-100 column. Columns were prepared using a 1ml syringe, plugged with a small ball of glass wool. TE buffer was passed through the column, followed by Sephadex G-100 slurry (Sigma Aldrich). Care was taken not to introduce air bubbles. Sephadex was added until there was no further settling.
5. The column was placed into a 15ml sterilin tube and centrifuged at 1000rpm for 5 minutes at room temperature. The fluid collected in the sterilin tube was discarded.
6. 100µl of TE was added to the column and step 5 repeated.
7. The column was placed into a clean sterilin tube. 80µl of TE was added to the labelled probe mixture and the new volume pipetted into the column.
8. This was centrifuged at 1000rpm for 5 minutes at room temperature. The column was discarded and the fluid collected in the sterilin tube, containing the labelled probe, was transferred to a 1.5ml microcentrifuge tube.
9. The labelled probe was used immediately or stored at -20°C.

2.7 Filter Hybridisation

Filter hybridisation was carried out in a hybridisation oven (Biometra). The hybridisation chambers were bottles that were clamped in a rotisserie and continuously rotated.

1. The filter was soaked in 2x SSC, placed on an appropriately sized piece of hybridisation mesh and positioned in a hybridisation bottle containing 5-10ml of 2x SSC. The bottle was placed in the rotisserie in the oven at 65°C.

2. The SSC was poured from the bottle and replaced with hybridisation buffer, 10ml/large bottle and 5ml/small bottle. The hybridisation buffer contained salmon sperm DNA (Sigma Aldrich) at a concentration of 10mg/ml, denatured by boiling for 7 minutes and quenched on ice for 3 minutes. The bottle was returned to the oven and the filter was prehybridised at 65°C for 1-12 hours. For probes containing multiple repeat sequences, human placental DNA, denatured by boiling for 7 minutes and quenched on ice for 3 minutes, was added to the hybridisation buffer at a concentration of 50µg/ml.
3. The labelled probe was denatured by boiling for 7 minutes, quenched on ice for 3 minutes and added to 5ml or 10ml (depending on bottle size) of fresh hybridisation buffer.
4. The prehybridisation buffer was poured from the bottle and replaced with the labelled probe and fresh hybridisation buffer. The bottle was returned to the oven and the filter hybridised at 65°C for 16-24 hours.

2.8 Filter Washing

1. The filter was removed from the bottle and rinsed by immersing in 1 litre of 4x SSC, 0.1% SDS.
2. The filter was transferred to 1l of 1x SSC, 0.2% SDS and incubated at 65°C for 30 minutes.
3. The filter was transferred to a second hot wash of 1l of 0.2x SSC, 0.2% SDS and incubated at 65°C for 5-30 minutes. The length of time incubated was determined by regular monitoring of the filter. The filter was periodically removed from the wash solution and monitored using a Geiger counter. When a background reading of 2-5 counts per second was obtained, the filters were regarded as adequately washed and were removed from the solution.
4. The filter was wrapped in Saranwrap (Fahrenheit) and placed between two sheets of X-ray film (Fuji, Genetic Research Instrumentation, Essex, UK), between intensifying screens at -70°C, in order to obtain an autoradiographic image. The top sheet of film was developed after 1-3 days and the lower sheet

of film developed 1-14 days later, depending on signal intensity on the first exposure.

2.9 Removal of the probe from the filter

Filters were kept wrapped in Saranwrap at -70°C to prevent dehydration. Prior to rehybridisation the filter was stripped of the probe from the previous hybridisation.

1. The filter was washed in 0.4M NaOH for 20 minutes at 45°C.
2. The filter was then transferred to a neutralising solution, 0.1x SSC, 0.1% SDS and 0.2M Tris HCl pH7.5, for 15 minutes at 45°C.

2.10 Autoradiography

Autoradiographic signal intensities were quantitated using an LKB 2222-020 Ultrascan XL Laser Densitometer, according to the manufacturer's instructions.

2.11 Preparation of single-stranded templates for DNA sequencing

2.11.1 Cloning of the insert from the plasmid into M13 phage

2.11.1.1 Preparation of insert and vector DNA

1. Approximately 1µg of plasmid DNA was digested with the appropriate restriction enzymes to release the insert. The enzymes were selected because they were present in the multiple cloning site of both the plasmid and the M13 vector strain used. The plasmid was digested in a total volume of 50µl, with 5µl of 10x buffer, 5µl of 0.1% bovine serum albumin (BSA), 5µl of 0.1% Triton X-100 (Sigma Aldrich) and 40 units of each restriction enzyme. Sterile distilled water was added to make the total volume 50µl.
2. The digest was incubated at 37°C for at least 2 hours (up to overnight).

3. Following digestion the plasmid DNA was electrophoresed on a 1% LMT (low melting temperature) agarose gel, prepared with 1x TAE buffer.
4. The plasmid insert was excised and purified using Wizard® columns. It was stored at -20°C and used as required.
5. Approximately 5µg of M13 vector was digested with the appropriate restriction enzymes as described, and run on a 1% LMT agarose gel. The digested M13 vector was excised and purified using Wizard® columns and stored at -70°C and used as required.

2.11.1.2 Ligation of the insert into the M13 vector

1. Two ligation reactions were set up for each insert to be cloned, in order to create two insert:vector ratios. Each ligation reaction was set up in a total volume of 10µl, with 1µl of ligation buffer, 1µl (~20ng) of M13 cut with the appropriate restriction enzymes, 0.5µl of T4 DNA ligase (Roche Products) and insert DNA to give an insert:vector ratio of 2:1 or 0.5:1. Sterile distilled water was added to make the total volume 10µl.
2. Ligation reactions were incubated at 16°C for at least 1 hour and up to overnight.
3. Prior to transformation, the required volume (2-10µl) of each ligation reaction was transferred to a sterile 14ml transformation tube (Falcon, Becton Dickinson, Oxford, UK).
4. A ligation control was also set up in order to check that the vector had been correctly prepared, this was as described, but without the insert DNA.

2.11.2 Preparation of competent *E.coli* for transformation

The preparation of competent *E.coli* is based on the method of Hanahan *et al.*, 1991.

1. JM101 cells were stored in glycerol at -70°C; prior to use the cells were streaked onto a minimal media plate and incubated overnight at 37°C. The minimal plate was then stored at 4°C, the cells were viable for several weeks.

2. A colony was picked from the minimal plate, streaked onto a SOB plate and incubated at 37°C overnight.
3. A few colonies were lifted from the SOB plate and used to inoculate 50ml of sterile SOB media in a sterile 500ml conical flask.
4. The inoculated media was incubated at 37°C with shaking, until the OD_{550nm} reached 0.3-0.4.
5. 5mls of the culture was taken for the lawn. It was added to 5mls of fresh SOB media and incubated further at 37°C without shaking to give a dense culture.
6. The remaining culture was divided between 2 pre-cooled 50ml conical tubes and cooled on ice for 15 minutes.
7. The culture was centrifuged at 2500rpm for 10 minutes at 4°C. The supernatant was poured off and the tubes inverted and tapped sharply on tissue, in order to remove as much SOB as possible.
8. The cell pellets were resuspended in 1/3 volume TFB, that is, 1/3 of the original culture volume (for a 50ml culture this was 16.6ml, 8.3ml per tube). The cells were resuspended by gentle swirling of the TFB over the pellet and incubated on ice for 10 minutes.
9. The cells were centrifuged for 10 minutes at 4°C and the supernatant removed as previously. The cell pellets were resuspended in 1/12.5 volume TFB (i.e. 4ml total, 2ml per tube) and the tubes pooled.
10. 28µl of nitrogen-purged, top grade DMSO (dimethyl sulfoxide) (Sigma Aldrich) per 10ml of original starting culture (i.e., 140µl for a 50ml culture) was added and the cells incubated on ice for 10 minutes.
11. 28µl of 2.2M DTT (Sigma Aldrich) in 10mM KOAc pH6.2 per 10ml of initial culture was added and the cells incubated on ice for 10 minutes.
12. Step 10 was repeated with an incubation of 5 minutes instead of 10 minutes. The cells were now competent.

2.11.3 Transformation of competent cells with ligated M13

1. 200µl of competent cells were added to each 14ml transformation tube containing 2-10µl of a ligation reaction (section **2.11.1.2**) and incubated on ice for 40 minutes.
2. The cells were heat-shocked at 42°C for 2 minutes and then returned to ice.
3. JM101 lawn cells (section **2.12.2**, step 5) were mixed with 2x YT soft agar (melted and held at 45°C) in the proportions 2ml of lawn cells per 35ml of soft agar. 945µl of 2% X-gal in dimethylformamide and 450µl of 2% IPTG (in water) per 35ml of agar were added.
4. 4ml of the agar mix was added to each tube of heat shocked cells. This was mixed by rolling and poured immediately onto 100mm 2x YT agar plates.
5. The plates were allowed to set at room temperature, inverted and incubated at 37°C overnight.
6. Control transformations on duplicate plates were also carried out with; cut vector at 10ng/plate, ligation control at 10ng/plate and uncut vector at 1ng/plate.

2.11.4 Preparation of single-stranded templates

1. 100ml of 2x YT was inoculated with 1ml of an overnight standing culture of JM101 cells in 2x YT.
2. This was dispensed in 1.5ml aliquots into sterile 50ml conical tubes.
3. Each tube was inoculated with a colourless plaque using sterile 1-10µl pipette tips. The pipette tip was used to stab the plaque and was then dropped into the tube. It was removed after a few minutes.
4. The tubes were shaken at 37°C for 6 hours.
5. The culture was transferred into a 1.5ml microcentrifuge tube and centrifuged at 13000 rpm for 5 minutes at room temperature.
6. The supernatant was decanted into a fresh tube containing 200µl 20% PEG/2.5M NaCl. It was mixed well and incubated at room temperature for 20 minutes.

7. The tubes were centrifuged at 13000 rpm for 5 minutes at room temperature and the supernatant discarded.
8. They were centrifuged for a further 2 minutes and all traces of PEG carefully removed using a drawn out Pasteur pipette.
9. The viral pellet was resuspended in 200µl of sterile distilled water.
10. When the pellet was fully dissolved, 200µl of phenol was added, the mix vortexed for 5-10 minutes and then centrifuged for 5 minutes.
11. The upper aqueous layer was removed to a fresh microfuge tube.
12. 200µl of chloroform was added, the mix vortexed for 2 minutes and centrifuged for 2 minutes.
13. Steps 11 and 12 were repeated.
14. The upper aqueous layer was removed to a fresh microfuge tube and the single stranded template precipitated by adding 20µl of 3M sodium acetate and 600µl of 100% ethanol and incubating at -20°C overnight.
15. The template was pelleted by centrifuging at 13000rpm for 10-15 minutes. The supernatant was discarded and the pellet washed in 70% ethanol.
16. The pellet was dried, redissolved in 25µl of sterile distilled water and stored at -20°C

2.12 Automated fluorescent dye sequencing

The sequencing reactions were carried out using the Cy5 Autoread sequencing kit (Amersham Pharmacia Biotech) which incorporates the fluorescent dye Cy5 Amidite and all sequencing reactions were run on the ALF*express* automated DNA sequencer (Amersham Pharmacia Biotech). Two slightly different methods were used for the sequencing of single-stranded templates, the first used a Cy5 labelled primer and the second used Cy5 dATP for internal labelling. All reagents for the sequencing reactions, including Cy5 labelled M13 universal and reverse primers, were provided in the Autoread sequencing kit, with the exception of the Cy5 dATP (Amersham Pharmacia Biotech) which was purchased separately.

2.12.1 Sequencing reactions using a Cy5 labelled primer

2.12.1.1 Annealing of the primer to a single-stranded template

1. Using sterile distilled water, the concentration of the template was adjusted so that 13µl contained 1-2µg of DNA.
2. The following were added to a 0.6ml thin-walled tube:

Template DNA	13µl
Cyanine-labelled Primer (2-10 pmol)	2µl
Annealing buffer	2µl
Total volume	17µl

3. The tube was vortexed gently, centrifuged briefly and incubated at 60°C for 10 minutes.
4. The tube was incubated at room temperature for at least 10 minutes and then centrifuged briefly.
5. 1µl of extension buffer was added to the annealing reaction and the sequencing reactions performed immediately.

2.12.1.2 Sequencing reactions

1. During the incubation steps of primer annealing, the sequencing mixes and T7 DNA polymerase were prepared for sequencing. Four wells of a microtitre plate were labelled 'A', 'C', 'G' and 'T' respectively and 2.5µl of the 'A' mix, 'C' mix, 'G' mix and 'T' mix were pipetted into the appropriate well. The dispensed sequencing mixes were stored on ice until required and just prior to use the microtitre plate was placed on the hot block and the sequencing mixes warmed to 37°C. The required amount of T7 DNA polymerase was diluted with enzyme dilution buffer to the appropriate concentration, 3 units/ 2µl for single-stranded sequencing.

2. 2µl of diluted T7 polymerase was added to the annealing reaction and mixed thoroughly with a pipette tip. 4.5µl was immediately pipetted into each of the prewarmed sequencing mixes.
3. The reactions were incubated at 37°C for 5 minutes exactly.
4. 5µl of stop solution was added to each reaction and mixed gently using a pipette tip.
5. The reactions were stored on ice (or at -20°C if the gel was to be run at a later date), until the sequencing gel was ready for loading. Just prior to loading the reactions were heated to 85-90°C for 2-3 minutes and then quenched on ice. 6µl was loaded into the appropriate wells of a sequencing gel.

2.12.1.3 Annealing of the primer to a double-stranded template

1. Using sterile distilled water, the concentration of the template was adjusted so that 16µl contained 5-10µg of DNA.
2. The following were added to a 0.6ml thin walled tube;

Template DNA	16µl
2 M NaOH	8µl
Distilled Water	16µl
Total volume	40µl

3. The tube was vortexed gently, centrifuged briefly and incubated at room temperature for 10 minutes.
4. 7µl of 3M sodium acetate (pH 4.8) and 4µl of dH_2O were added.
5. 120µl of 100% ethanol was added, mixed, and placed on dry ice for 15 minutes. (Alternatively, the tubes were incubated at -70°C for ≥30 minutes). The precipitated DNA was collected by centrifugation for 15 minutes. The supernatant was carefully removed and discarded, and the pellet rinsed with

70% ethanol, recentrifuged for 10 minutes, and the supernatant carefully removed. The pellet was dried at 37°C for 5 minutes.

6. The pellet was resuspended in 10µl of distilled water.
7. The following were added to the tube containing the resuspended template;

Template DNA	10µl
Cyanine-labelled Primer (4-10 pmol)	2µl
Annealing Buffer	2µl
Total Volume	14µl

8. The tube was vortexed gently, centrifuged briefly and the annealing reaction preheated at 65°C for 5 minutes. The reaction was immediately placed at 37°C and incubated for 10 minutes. The tube was removed and placed at room temperature for at least 10 minutes, then centrifuged briefly.
9. 1µl of Extension Buffer and 3µl of DMSO were added and the sequencing reactions performed immediately.

2.12.1.4 Sequencing reactions

1. The sequencing mixes and T7 DNA polymerase were prepared for sequencing, exactly as described above (section **2.12.1.2**), with the exception that the enzyme was diluted to a concentration of 6-8 units/ µl for double-stranded sequencing.
2. 2µl of diluted T7 polymerase was added to the annealing reaction and mixed thoroughly with a pipette tip. 4.5µl was immediately pipetted into each of the prewarmed sequencing mixes.
3. The reactions were incubated at 37°C for 5 minutes exactly.
4. 5µl of stop solution was added to each reaction and mixed gently using a pipette tip.

5. The reactions were stored on ice (or at -20°C if the gel was to be run at a later date), until the sequencing gel was ready for loading. Just prior to loading the reactions were heated to 85-90°C for 2-3 minutes and then quenched on ice. 6µl was loaded into the appropriate wells of a sequencing gel.

2.12.2 Sequencing reactions using a Cy5 dATP internal label

This protocol enables the use of unlabelled primers.

2.12.2.1 Annealing of the primer to a single-stranded template

1. The template was adjusted so that 12µl contained 1-2µg of DNA.
2. The following were added to a 0.6ml thin-walled tube;

Template DNA	12µl
Primer (4-400pmol)	2µl
Annealing buffer	2µl
Total volume	16µl

The primers used were custom made commercially.

3. The tube was vortexed gently, centrifuged briefly and incubated at 65°C for 10 minutes.
4. The tube was allowed to cool to less than 30°C over a period of 45 minutes and then centrifuged briefly. The sequencing reactions were performed immediately.

2.12.2.2 Sequencing reactions

1. During the slow cool of the annealing reactions the sequencing mixes and T7 DNA polymerase were prepared for sequencing, exactly as described above (section **2.12.1.2**), with the exception that 3µl of each of 'A' mix, 'C' mix, 'G' mix and 'T' mix were dispensed into the appropriate wells of the microtitre plate.

2. 1µl of Cy5 dATP labelling mix was added to the annealed template and mixed with a pipette tip.
3. 2µl of diluted T7 polymerase was added to each reaction and they were incubated at room temperature for exactly 5 minutes, (the addition of T7 polymerase was staggered to allow time for pipetting at the next stage).
4. When the incubation was almost complete, the dispensed sequencing mixes were pre-warmed to 37°C for at least 1 minute.
5. 1µl of extension buffer was added to each reaction and 4.5µl was immediately added to each of the pre-warmed sequencing mixes.
6. The reactions were incubated for 5 minutes.
7. 5µl of stop solution was added to each reaction and mixed by gentle agitation.
8. Immediately prior to loading the sequencing gel, the reactions were heated to 85°C-90°C for 2-3 minutes and then quenched on ice. 6-8µl of each reaction was loaded into the appropriated wells of a sequencing gel.

2.12.2.3 Annealing of the primer to a double-stranded template

The standard annealing of primer to double-stranded template was prepared, exactly as described above (section **2.12.1.3**), with the exception that the pellet was resuspended in 12µl of distilled water.

2.12.2.4 Sequencing reactions

The sequencing mixes and T7 DNA polymerase were prepared for sequencing, exactly as described above (section **2.12.2.2**), with the exception that the enzyme was diluted to a concentration of 6-8 units/µl for double-stranded sequencing; 3.5µl of DMSO was added to each reaction; 5.4µl immediately added to each of the pre-warmed sequencing mixes, and 6µl of Stop Solution was added to each reaction.

2.12.3 Preparation of the sequencing gel plates

The gel plates were cleaned using Kimwipe tissues (Kimberley Clark, Merck, Hertfordshire, UK), as these tissues are free of dyes that may cause interference when excited by the laser.

1. The gel plates were cleaned using a soft brush and a detergent, which did not fluoresce (Synerphonic N, Merck) and rinsed thoroughly with distilled water. The spacers and comb were rinsed in distilled water only.
2. Each item was wiped dry using lint-free tissues.
3. The two gel plates were cleaned again with lint-free tissue soaked in sterile distilled water, by wiping from the bottom of the plate to the top. This was repeated with 100% ethanol.
4. The top 1-2cm of each plate was wiped with diluted bind silane (Amersham Pharmacia Biotech) solution. This was left to dry for a few seconds and then polished with a lint-free tissue.
5. The plates were polished with ethanol, from bottom to top, taking care not to spread bind silane elsewhere on the plates.
6. The spacers were wiped with 100% ethanol and positioned on the thermoplate (bottom plate), and gentle pressure was applied to secure them to the silicone rubber seals. The top plate was lowered into position on the thermoplate and the two plates clamped together, using specially designed clamps.
7. The comb was wiped with 100% ethanol and placed in position at the top of the plates, ensuring it rested flush with the left hand side of the gel cassette.
8. The apparatus was levelled to ensure even distribution of the gel.

2.12.4 Preparation of the gel

All reagents used in the preparation of the gel were ALF grade (Amersham Pharmacia Biotech).

1. The following reagents were mixed together in a clean 250ml conical flask;

ALF grade urea	27g
Long Ranger polyacrylamide gel mix	9mls
(JT Baker, London, UK)	
10x TBE (ALF grade)	11.25mls
Sterile distilled water	up to 75mls

2. The mix was filter sterilised through a 0.22µ Falcon bottle top filter (Falcon, Becton Dickinson).
3. 37.5µl of temed (Sigma Aldrich) and 375µl of 10% ammonium persulphate (ALF grade) (Amersham Pharmacia Biotech) were added to the gel mix, it was mixed and poured into a 50ml syringe (with the plunger removed).
4. The plunger was placed back in the syringe and the gel solution was applied in an even motion, back and forth along the lower lip of the glass plate. Capillary action drew the solution front upward to fill the entire space between the plates.
5. The gel was left to polymerise for 3 hours.

2.12.5 Loading the gel

1. The gel cassette was placed in the ALF*express* and the top and bottom chambers filled with 0.6x TBE buffer. It was correctly linked to the cooling system.
2. The comb was removed and the wells flushed out with buffer using a needle and syringe.
3. The ALF*express* was programmed with the appropriate running details and the laser and temperature left to equilibrate.
4. The denatured sequencing reactions were loaded into the appropriate wells of the gel, the electrodes connected and the machine started.

2.13 RACE PCR

RACE PCR was carried out using Marathon-Ready™ cDNAs (Clontech, Basingstoke, UK).

1. Gene-Specific Primers (GSPs) were designed: 23-28 nucleotides in length, had a GC content of 50-70%, and a T_m of at least 65°C. The GSPs should be designed approximately 100-200bp towards the end of the 5' and/or 3' end of the sequence depending on whether 5' and/or 3' RACE is to be performed.
2. The following reagents were added to a 0.6ml thin walled tube;

Marathon-Ready™ cDNA	5µl
AP1 primer (10µM) (Clontech)	1µl
GSP1 (10µM)	1µl
Master Mix	43µl
Total volume	50µl

For the positive control, the Marathon-Ready™ cDNA was replaced with 5µl of G3PDH cDNA, and the GSP replaced with 1µl of the Control 5' and/or 3' G3PDH primer. For the negative control, the Marathon-Ready™ cDNA was replaced with 5µl distilled water.

3. The PCR MasterMix was prepared from the following reagents;

10x PCR reaction buffer	5µl
dNTP mix (2mM)	5µl
Ampli*Taq* Gold™ DNA Polymerase (2.5U)	0.5µl
Distilled water	32.5µl
Total volume	43µl

4. The Tm of the GSP determined whether a 2-step (Touchdown) PCR amplification, or a standard 3-step PCR amplification was carried out. The thermal profile selected for each GSP primer was essentially as shown below, but the optimal annealing temperature and program was substituted for the annealing temperature shown.

4a. Touchdown thermal profile:

Pre-Incubation/Activation Step	95°C for 10 minutes
5 cycles of;	94°C denaturation for 30 seconds 72°C primer annealing/extension for 4 minutes
5 cycles of;	94°C denaturation for 30 seconds 70°C primer annealing/extension for 4 minutes
20 cycles of; (25 cycles in nested reaction – see 5).	94°C denaturation for 30 seconds 68°C primer annealing/extension for 4 minutes
Final extension	68°C for 7 minutes
Hold temperature	4°C

4b. 3-step thermal profile:

Pre-Incubation /Activation Step	95°C for 10 minutes
30 cycles of, (35 cycles in nested reaction-see 5.)	94°C denaturation for 30 seconds 65°C primer annealing for 30 seconds 72°C primer extension for 4 minutes
Final extension	72°C for 7 minutes
Hold temperature	4°C

5. 3µl of first-round PCR product was removed from the tubes and added to a new sterile 0.6ml thin walled tube. For each 50µl nested PCR reaction, the following reagents were added to the first-round PCR product;

First-round PCR product	3µl
AP2 Primer (10µM) (Clontech)	1µl
GSP2 Primer (10µM)	1µl
Master Mix	45µl
Total volume	50µl

6. The tubes were mixed gently and centrifuged briefly to bring the contents to the bottom of the tube.
7. Approximately 60µl of mineral oil was added to each tube, to prevent any evaporation, and the tubes centrifuged briefly again.
8. The Tubes were placed on the thermal cycler (Biometra Trio Thermoblock) and subjected to the appropriate thermal profile. Each GSP2 primer, like GSP1, had an optimal annealing temperature, which determined whether a 2-step (Touchdown) PCR amplification, or a standard 3-step PCR amplification was carried out (see 4a and 4b).
9. The mineral oil was removed from the tubes and 5µl was run on a 1% mini agarose gel.
10. The gel was stained with ethidium bromide and visualised on a UV transilluminator. If a single band of size >100bp was observed, more often in each tissue, then the remainder of the PCR product was purified. Purification was achieved by the use of Wizard® PCR preps as previously described in section **2.5.1.3**.

2.14 Subcloning of RACE PCR products for direct sequencing

The purified PCR products were cloned into the pGEM®-T Easy Vector System II (Promega).

2.14.1 A-Tailing of RACE PCR products

For each A-Tailing reaction, the following were added to a sterile 0.6ml thin-walled tube;

10x reaction buffer	1µl
50mM $MgCl_2$	0.5µl
dATP (1:100)	2µl
Purified PCR product	2-6µl
Taq polymerase (2.5U)	0.5µl
Sterile distilled water	up to 10µl

and incubated at 70°C for 30 minutes.

2.14.2 Ligation of RACE PCR products into the vector

Two insert:vector ratios (2:1 and 1:2) for each A-tailed PCR product were used when ligating into the pGEM®-T Easy Vector. For each PCR product ratio, including positive and background controls, the following were added to a 0.6ml sterile thin walled tube;

2x Rapid Ligation Buffer	5µl
pGEM®-T Easy Vector (50ng)	1µl
PCR product	2µl/0.5µl
Control Insert DNA (positive control only)	1µl
T4 DNA Ligase (3 Weiss units/µl)	1µl
Sterile distilled water	up to 10µl

and ligated at room temperature for ≥ 1 hour or overnight at 4°C.

2.14.3 Transformation of competent cells

50µl of JM109 High Efficiency Competent Cells were added to each 10µl ligation reaction and incubated on ice for 20 minutes. The cells were then heat-shocked for 50 seconds at 42°C and immediately returned to ice for 2 minutes. 950µl of room temperature SOC medium was added to each transformation tube, and incubated for 1.5 hours at 37°C with shaking (~150rpm). Following incubation, 200µl of each transformation culture was plated onto LB/ampicillin/IPTG/X-Gal plates. Alternatively, to increase the number of colonies, the cells were pelleted by centrifugation at 1,000 x *g* for 10 minutes, resuspended in 200µl of SOC medium and plated out. The plates were then incubated overnight (16-24 hours) at 37°C.

2.15 Localisation of known genes and novel cDNAs to the YAC contig

1. Gene specific primer pairs were designed for each known gene or novel cDNA. The primers were approximately 50% GC and at least 19 bases in length. Additionally, the primers contained either a G or C residue as the last 3'-base, and did not have any regions that could self-anneal or form "hair pin" loops. The primers were dissolved in distilled water to a final concentration of 100pm/µl.
2. For each 100µl PCR reaction the following were added to a sterile 0.6ml thin walled tube;

Sterile distilled water	up to 100µl
10x reaction buffer	10µl
50mM $MgCl_2$	6µl
dNTP mix (1.25mM)	8µl
Primer 1 (100pmol/µl)	1µl
Primer 2 (100pmol/µl)	1µl
YAC template DNA (≈200ng)	1µl
Taq polymerase (2.0-2.5 units)	0.5µl

The buffer, $MgCl_2$, and *Taq* polymerase were manufactured by Bioline. The dNTP mix (Amersham Pharmacia Biotech) was prepared by mixing 12.5µl of each of the 100mM dNTPs with 950µl of sterile distilled water and storing in 1ml aliquots. The volume of 50mM $MgCl_2$ stated was a standard starting point, if necessary a $MgCl_2$ titration was performed to determine the optimal final concentration. The template was YAC DNA from the YAC contig spanning the critical region of the 5q- syndrome (Kostrzewa *et al.*, 1998).

3. The tubes were mixed gently and centrifuged briefly to bring the contents to the bottom of the tube.
4. Approximately 60µl of mineral oil was added to each tube, to prevent any evaporation, and the tubes centrifuged briefly again.
5. The tubes were placed on the thermal cycler (Biometra Trio Thermoblock) and subjected to the appropriate thermal profile. Each primer pair had an optimal annealing temperature, and the thermal profile selected for the primer set was as shown below:

Initial denaturation	95°C for 5 minutes
35 cycles of;	95°C denaturation for 30 seconds
	60°C primer annealing for 30 seconds
	72°C primer extension for 1 minute
Final extension	72°C for 8 minutes
Hold temperature	4°C

6. The mineral oil was removed from the tubes and 5µl was run on a 1.5% mini agarose gel.
7. The gel was stained in ethidium bromide and visualised on a UV transilluminator.

2.16 Reverse transcriptase PCR (RT-PCR) analysis

RT-PCR analysis was carried out using the R*everse*--iT™ One-step PCR kit (ABgene, Surrey, UK).

1. RT-PCR primer pairs were designed flanking the coding region of known genes and novel cDNAs. The primers were approximately 50% GC and at least 19 bases in length. Additionally, the primers contained either a G or C residue as the last 3'-base, and did not have any regions that could self-anneal or form "hair pin" loops. The primers were dissolved in RNase-free water to a final concentration of 100pm/µl.
2. The following were added to a 0.5ml RNase-free microfuge tube;

2x Reddy-Load Master Mix	25µl
Sense primer (10µM)	1µl
Anti-sense primer (10µM)	1µl
Reverse Transcriptase Blend	1µl
RNA template (1µg/µl)	1µl
RNase-Free Water	21µl
Total volume	50µl

For the positive control, 1µl of MS2 positive control template and 1µl of MS1 and MS2 primers was substituted for the experimental RNA template and primers, respectively

3. The tubes were mixed gently and centrifuged briefly to bring the contents to the bottom of the tube.
4. The tubes were placed on the thermal cycler (Hybaid PCR sprint, Hybaid, Middlesex, UK) and subjected to the appropriate thermal profile. Each primer pair had an optimal annealing temperature and the thermal profile selected for the primer set was as shown below:

1st strand synthesis	47°C for 30 minutes
RTase inactivation and	
initial denaturation	94°C for 2 minutes
40 cycles of;	94°C denaturation for 20 seconds
	63°C primer annealing for 30 seconds
	72°C primer extension for 1 minute
Final extension	72°C for 5 minutes
Hold temperature	4°C

2.16.1 Purification and quantification of RT-PCR products

1. 1µl of purified template, 1µl of Bromophenol Blue loading dye, and 3µl of distilled water to a final volume of 5µl was added to a 0.5ml microcentrifuge tube.
2. The tubes were mixed gently and centrifuged briefly to bring the contents to the bottom of the tube.
3. All 5µl was run on a 1% Low EEO Agarose gel in 1x TAE buffer at 100V for approximately 40 minutes. The purified product should be seen as a high quality, single band.
4. A small aliquot of the reaction (5µl) was run on a 1% mini agarose gel.
5. The gel was stained in ethidium bromide and visualised on a UV transilluminator.

2.17 CD34+ expression by RT-PCR

Total RNA was obtained from CD34+ cells with the Total RNA Isolation Kit (AMS Biotechnology (Europe) Ltd., Oxfordshire, UK). RT-PCR was carried out using the R*everse*--iT™ One-step PCR kit (ABgene) as previously described (**2.16 steps 2-4**).

1. RT-PCR primer pairs were designed flanking the coding regions of each novel cDNA/gene. The primers were dissolved in RNase-free water to a final concentration of 100pm/µl.

2. $CD34^+$ RNA (0.1µg/ml) was used as the template in the RT-PCR reaction. For the positive control, 1µl of a Total RNA mononuclear fraction from a healthy individual was substituted for the $CD34^+$ RNA template.

2.18 Cycle sequencing on the ALF*express* automated sequencer

All cycle sequencing reactions were carried out using the Thermo Sequenase™ Cy™5 Dye Terminator Kit (Amersham Pharmacia Biotech), for use on the Alf*express* automated sequencer (Amersham Pharmacia Biotech).

2.18.1 Preparation of dNTP/Cy5 ddNTP mixes

All mixes were prepared on ice.

1. 4 tubes were labelled "A Mix", "C Mix", "G Mix", and "T Mix respectively for preparation of the dye terminator mixes.
2. 4µl of 1.1mM dNTP, 2µl of the appropriate Cy5 ddNTP, and 16µl of distilled water to a final volume of 22µl was added to each "Mix". These volumes are sufficient for 10 sequencing reactions.
3. The mixes were vortexed and centrifuge briefly to collect the contents at the bottom of each tube.
4. The dNTP/Cy5 ddNTP mixes were stored on ice, in the dark, until needed.

2.18.2 Sequencing reactions

All sequencing reactions were prepared on ice.

1. Four sterile 0.6ml thin walled tubes "A", "C", "G", and "T" were labelled for each template (1-10), respectively.
2. 2µl of the "A Mix", "C Mix", "G Mix", and "T Mix" (prepared in 4.2.15.4) were dispensed into the corresponding tubes "A", "C", "G", and "T" respectively.

3. A Master Mix was prepared for each template (set of 4 reactions; A, C, G, and T), by combining the following in a 0.5ml microcentrifuge tube;

Template DNA#	1.0-20.5µl
Primer (4pmol)*	2µl
Reaction buffer	3.5µl
Thermo Sequenase DNA polymerase (10U/µl)	1µl
Distilled water	to a final volume of 27µl

Generally, 1-4µl of template was used in the sequencing reaction if a clean, bright band was seen on an agarose gel. 10µl was used if the band was faint. Anything over 10µl did not give a sufficient quality sequence.

* Nested primers were designed a few base pairs internally to the amplification primers for the sequencing reactions as they produced a better sequence run due to the added specificity. The primers were diluted in distilled water to a final concentration of 2pmol/µl.

4. The tubes were vortexed and then centrifuged briefly to bring the contents to the bottom of the tube. 6µl of each template Master Mix was aliquoted into each of its "A", "C", "G", and "T" mixes. The reactions were mixed by gently pipetting up and down. The reactions were overlayed with 10µl of mineral oil if a non-heated lid thermal cycler was used.
5. The tubes were placed on the thermal cycler and subjected to the appropriate thermal profile. Each primer had an optimal annealing temperature, and the thermal profile selected for both primers was as shown below:

30 cycles of;	95°C	for 30 seconds
	61°C	primer annealing for 30 seconds
	72°C	primer extension for 1 minute 20 seconds
	4°C	hold

2.18.3 Precipitation of sequencing reactions

Precipitation of all sequencing reactions was carried out on ice.

1. If a non-heated lid thermal cycler was used, each sequencing reaction was transferred from underneath the mineral oil to a new 0.5ml microcentrifuge tube.
2. 2µl of 7.5M ammonium acetate was added directly to each reaction.
3. 2µl of glycogen solution was added to each tube.
4. 30µl of ice-cold absolute ethanol was added to each tube.
5. The samples were vortexed and incubated on ice for 20 minutes. Alternatively, they were incubated overnight at -20°C.
6. The reactions were centrifuged at full speed (10,000-16,000xg) in a microcentrifuge for 15 minutes at 4°C to pellet the DNA.
7. The supernatant was removed, making sure the pellet was not touched. 200µl of ice-cold 70% ethanol was added to each pellet.
8. The reactions were centrifuged at full speed for 5 minutes at 4°C.
9. The supernatant was removed carefully, firstly with a 20-200µl pipette tip, then with a 0.5-10µl pipette tip. Any ethanol remaining in the tube was removed with a clean, lint and dye free tissue, making sure the pellet was not disturbed. The pellets were left to dry for no longer than 10 minutes, as overdrying made the pellets difficult to resuspend. In addition, the pellet turned from being bright white to transparent if it was left to dry for too long, making it less visible.
10. The pellet was resuspended in 8µl of Stop solution by pipetting slowly up and down.

2.18.4 Preparation of the polyacrylamide gel and loading of samples

1. The glass gel plates were cleaned twice with distilled water. Kimwipe tissues were used at all times for cleaning the gel plates. Kimwipe tissues were used

because they contain no dye, and therefore prevent contaminating fluorescence.

2. The glass gel plates were cleaned twice with Absolute ethanol.
3. The top inch of both plates was wiped with Bind-Silane.
4. The glass gel plates were cleaned again with Absolute ethanol.
5. Two 0.5mm glass spacers were cleaned with Absolute ethanol and placed securely on the bottom Thermoplate (Amersham Pharmacia Biotech). The top plate was placed on top and secured with clips.
6. The 0.5mm comb was cleaned with Absolute ethanol and placed between the two plates.
7. ReproGel™ High Resolution gel mix (Amersham Pharmacia Biotech) Solution B was added to Solution A. The polyacrylamide was mixed by inverting the bottle 5 times. The gel mix was loaded directly onto the gel plates.
8. The gel plates were placed directly under the ReproSet™ (Amersham Pharmacia Biotech) and the gel mix left to polymerise for 10 minutes under the UV light.
9. The samples were denatured at 72°C for 3 minutes, then immediately placed on ice. If the samples were denatured at the normal 95°C, the signal from the larger dye-terminated fragments would be degraded.
10. The entire volume of each reaction (8µl) was loaded into the appropriate well of the sequencing gel and electrophoresed under the appropriate conditions.

2.18.5 Processing the sequence data on the ALF*express* automated sequencer

Each clone (1-10) was processed using the Extended Shift Function available in ALFwin™ Sequence Analyser 2.10 software (Amersham Pharmacia Biotech). This process overcomes the effect known as "smiling" that is produced when samples in the first and last lanes move more slowly than those in the middle lanes. The data was always processed after the run had completed. The software never processed the data as a post-run action.

2.19 Cycle sequencing on the ABI PRISM 3100 Genetic analyser

All cycle sequencing reactions were carried out using the ABI PRISM® BigDYE™ Terminator Cycle Sequencing Ready Reaction Kit v2.0 (Applied Biosystems UK, Warrington, Cheshire) for use on the ABI PRISM 3100 Genetic analyser (Applied Biosystems).

2.19.1 Preparation of sequencing reactions

Preparation of sequencing reactions was carried out at room temperature.

1. Sixteen sterile 0.6ml thin-walled tubes were labelled for each of the 8 templates: (1-8 for Forward primer; 9-16 for Reverse primer).
2. A Master Mix was prepared for each primer (9 reactions each) by combining the following in a 1.5ml microcentrifuge tube;

Template DNA*	1.0-2.0µl
Primer (3.2 pmol)	1µl
5x Sequencing buffer	2µl
Terminator Ready Reaction Mix	4µl
Distilled water	to a final volume of 20µl

* Generally, 1-2µl of a 1:10 dilution of the genomic DNA PCR product was used as the template in the sequencing reaction if a clean, bright band was seen on an agarose gel.

3. Each Master Mix was vortexed to mix. 19µl of Master Mix was added to 1µl of template. The reactions were vortexed and then centrifuged briefly to bring the contents to the bottom of the tube.

4. The tubes were placed on the Hybaid PCR Sprint thermal cycler (Hybaid Limited, Middlesex, UK) and subjected to the appropriate thermal profile:

25 cycles of;	96°C for 10 seconds
	50°C primer annealing for 5 seconds
	60°C primer extension for 4 minutes
	4°C hold

2.19.2 Sodium acetate precipitation of sequencing reactions

Precipitation of all sequencing reactions was carried out at room temperature.

1. 2µl of 3M sodium acetate (pH 4.6) was added directly to each reaction.
2. 50µl of 95% ethanol was added to each tube.
3. The samples were vortexed and then centrifuged briefly to bring the contents to the bottom of the tube.
4. The reactions were incubated at room temperature for 15 minutes.
5. The reactions were centrifuged at full speed (10,000-16,000 x g) in a microcentrifuge for 20 minutes to pellet the DNA.
6. The supernatant was removed, making sure the pellet was not touched. 250µl of 70% ethanol was added to each pellet.
7. The reactions were vortexed briefly and centrifuged at full speed for 5 minutes.
8. The supernatant was removed carefully, firstly with a 20-200µl pipette tip, then with a 0.5-10µl pipette tip. The pellets were left to dry for no longer than 20 minutes, as overdrying made the pellets difficult to resuspend.
9. 20µl of Template Suppression reagent (TSR) was added to each reaction. The tubes were vortexed and centrifuged briefly to bring the contents to the bottom of the tube.
10. The samples were denatured at 95°C for 2 minutes, then immediately chilled on ice.

2.19.3 Processing the sequence data on the ABI PRISM 3100 Genetic analyser

All sequence data was collected using the ABI PRISM Data Collection Software version 1.0, and processed using the ABI PRISM DNA Sequencing Analysis Software version 3.6 NT.

Chapter 3

Identification, localisation, cloning, and mutation analysis of novel gene *C5orf4*

3.1 Introduction

3.1.1 Identification of disease genes

3.1.2 Positional candidate gene approach

3.1.3 Chromosome 5 mapped genes

3.1.4 Isolating expressed sequence tags (ESTs) from the critical region of the 5q- syndrome

3.2 Materials and Methods

3.2.1 ESTs

3.2.2 I.M.A.G.E. cDNA clones

3.2.3 Samples

3.2.4 Gene dosage analysis

3.2.5 Northern analysis

3.2.6 Southern analysis

3.2.7 Direct sequencing

3.2.8 Overlapping cDNA clones

3.2.9 cDNA library screening

3.2.10 RACE PCR

3.2.11 Localisation to the YAC contig

3.2.12 Database analysis using the Genetics Computer Group (GCG) software package

3.2.12.1 FastA

3.2.12.2 BLAST

3.2.12.3 Frames

3.2.12.4 Translate

3.2.12.5 Motifs

3.2.12.6 GenBank submission

3.2.13 Mutation analysis of *C5orf4*

3.2.13.1 Samples

3.2.13.2 Reverse Transcriptase PCR (RT-PCR)

3.2.13.3 Cycle sequencing

3.2.14 Database analysis using the Genetics Computer Group (GCG) software package
3.2.14.1 BestFit analysis

3.3 Results
3.3.1 ESTs
3.3.2 I.M.A.G.E. cDNA clones
3.3.3 Gene dosage analysis
3.3.4 Northern analysis
3.3.5 Southern analysis
3.3.6 Direct sequencing
3.3.6.1 I.M.A.G.E. cDNA clone 469867
3.3.6.2 I.M.A.G.E. cDNA clone 296617
3.3.7 Overlapping cDNA clones
3.3.7.1 I.M.A.G.E. cDNA clone 982453
3.3.7.2 I.M.A.G.E. cDNA clone 209846
3.3.7.3 I.M.A.G.E. cDNA clone 280058
3.3.7.4 False positive I.M.A.G.E. cDNA clones 935769 and 251760
3.3.8 cDNA library screening
3.3.9 RACE PCR
3.3.10 I.M.A.G.E. cDNA clone 435297
3.3.11 Localisation to the YAC contig
3.3.12 Database analysis using the Genetics Computer Group (GCG) software package
3.3.12.1 FastA analysis
3.3.12.2 BlastX analysis
3.3.12.3 Frames analysis
3.3.12.4 Translation of the novel cDNA
3.3.12.5 Motif search
3.3.12.6 GenBank submission
3.3.13 Mutation analysis of *C5orf4*

3.4 Discussion
3.4.1 Molecular studies
3.4.2 Translation of *C5orf4*
3.4.3 Mutation analysis of *C5orf4*

3.5 Conclusion

3.1 Introduction

3.1.1 Identification of disease genes

The choice of strategy for identifying a disease gene is dependent on the resources (animal models, chromosomal abnormalities, clone libraries, etc.) available, and on how much is known about the pathogenesis of the disease. Several strategies initially identify a number of candidate genes, which then are tested individually for evidence that implicates them as the disease locus. A number of different methods have been used to identify candidate genes, but mapping the disease to a specific sub-chromosomal localisation is generally the most productive first step.

3.1.2 Positional candidate gene approach

Once a disease has been mapped, it is now possible to use database searches to identify candidate genes. With increasingly more human genes being mapped to specific sub-chromosomal regions, positional candidate gene approaches are now dominating the field.

3.1.3 Chromosome 5 mapped genes

Human chromosome 5 contains an estimated 194 million bases, or approximately 6% of the human genome. A number of disease-linked genes have been mapped to chromosome 5. They include those for colorectal cancer, dwarfism, severe combined immunodeficiency, schizophrenia, basal cell carcinoma, deafness, atrial septal defect, asthma, and acute myelogenous leukaemia (HGP Information, April 2000 (http://www.ornl.gov/hgmis)). Several genes have been assigned in particular to 5q including many haematopoietic growth factors, for example the interleukin genes (Le Beau *et al.*, 1989), and the interferon regulatory factor 1 (*IRF1*) gene (Itoh *et al.*, 1991). Chromosome 5 is currently being sequenced by the Department of Energy's Joint Genome Institute in Walnut Creek, California as part

of the HGP. The draft sequence of chromosome 5, along with chromosomes 16 and 19 was released to the public on April 13, 2000. At this time it was believed that these three chromosomes contain an estimated 10,000-15,000 genes (HGP Information, April 2000). By February 2001, 24577kb was finished sequence (12.7%), represented by 152 contigs. The GeneMap of chromosome 5 from Entrez at NCBI
(http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/maps.cgi?ORG=hum&CHR=5) showed there were 331 known genes as of November 12, 2000, of which 47 map to the critical region of the 5q- syndrome at 5q31.3-q33.

3.1.4 Isolating expressed sequence tags (ESTs) from the critical region of the 5q- syndrome

Until the draft sequencing and annotation of the human genome was reported in February 2001, the most recent and advanced technique for the isolation of novel coding sequences has been the use of ESTs (Expressed Sequence Tags). ESTs are partial sequences, approximately 300-400bp, of the 3' and 5' ends of cDNAs. It is probable that the majority of the 30-40,000 genes in the human genome are now represented by one or more ESTs. On November 2, 2001 the number of public entries in the EST database db(EST)
(http://www.ncbi.nlm.nih.gov/dbEST/index.html) was 9,407,866, of which 3,876,441 were human. Moreover, there were over 1,000 separate entries comprising of EST clusters assigned to chromosome 5. ESTs representing novel genes can be accessed from the National Centre for Biotechnology Information (NCBI) website through The Human Gene Map
(http://www.ncbi.nlm.nih.gov/genome/guide/HsChr5.shtml), and UniGene
(http://www.ncbi.nlm.nih.gov/UniGene/Hs_DATA/ChromLists/Chr5.html). The former enables a specific sub-chromosomal region between specific markers to be accessed for ESTs represented as unidentified transcripts.

We selected ESTs for analysis that had been mapped to the region by two independent groups and contained sequences derived from cDNA libraries of haematological tissue, e.g. foetal liver spleen. The UniGene site has clusters that contain sequences that represent a unique gene. Each cluster contains a number of EST sequences with added information. We selected ESTs for analysis that either possessed a poly-adenylation signal; were novel; had similarity to known proteins of particular interest (after translation); contained a mapped sequence-tagged site (STS); or its clone source was a CGAP (Cancer Genome Anatomy Project) library. CGAP is an interdisciplinary program established and administered by the National Cancer Institute (NCI) to generate the information and technological tools needed to decipher the molecular anatomy of the cancer cell (CGAP homepage). This and other projects will form the basis for human genome research during the next few years, as the complete reference sequence becomes available to underpin the next phase of human biology and genetics.

Therefore, the EST database was used to identify novel cDNAs mapping to the critical region of gene loss, with the ultimate aim of isolating the putative tumour suppressor gene associated with the development of the 5q- syndrome.

3.2 Materials and Methods

3.2.1 ESTs

The Human Gene Map of chromosome 5 was accessed to identify ESTs assigned between the DNA markers D5S410 and D5S487 that span the critical region of the 5q- syndrome at 5q31.3-q33. One of the EST clusters identified revealed an unidentified transcript represented by 21 ESTs from the Soares foetal liver spleen 1NFLS library; the Soares melanocyte 2NbHM library; the Soares multiple sclerosis 2NbHMSP library; the Soares pregnant uterus NbHPU library; the Stratagene fibroblast (937212) library; and the Stratagene lung (937210) library. A transcript was selected that had no homology to any known genes and had ESTs from cDNA libraries of haematological origin.

Two ESTs were selected for further analysis based on their cDNA library source (Soares pregnant uterus NbHPU *Homo sapiens* library and the Soares foetal liver spleen 1NFLS library), and having no significant homology to any known human gene (i.e., representing novel genes).

3.2.2 I.M.A.G.E. cDNA clones

I.M.A.G.E. cDNA clones from which the two ESTs were derived were obtained from the Human Genome Mapping Project Resource Centre (HGMP-RC), Hinxton, Cambridge as stabs in agar. Single colonies were obtained by plating onto LB (Luria Bertani) ampicillin (50mg/ml) plates. A single colony was then inoculated into a 10ml LB culture containing ampicillin. Plasmid DNA was obtained using the QIAprep® Spin Miniprep Kit (QIAGEN). Both inserts were excised with the restriction enzymes *Not*I and *Eco*RI.

3.2.3 Samples

Six patients with the classical features of the 5q- syndrome, including a 5q deletion as the sole karyotypic abnormality were included in the study. Granulocyte and mononuclear cells were separated from 40mls of peripheral blood by ficoll gradient centrifugation (Boyum, 1984). The granulocytes showed a high level of purity (≥95%). Mononuclear cells (specifically T-lymphocytes) were isolated by erythrocyte rosetting and showed a purity of ≥90%. High molecular weight DNA was obtained from the fractionated blood leukocytes by Nucleon® extraction (Nucleon® Biosciences, Scotlab). Granulocyte DNA fractions from the peripheral blood of healthy individuals were used as controls. High molecular weight DNA was obtained from a human/mouse hybrid cell line with human chromosome 5 as its only human complement (GM11714, Coriell Cell Repositories, Camden, NJ).

3.2.4 Gene dosage analysis

Gene dosage analysis was used to confirm that the cDNA mapped to the critical region of the 5q- syndrome and to quantitatively assess the allelic loss of the I.M.A.G.E. cDNA clone. Gene dosage compares the hybridisation signal intensity from the gene of interest with the signal intensity from the gene on an uninvolved chromosome, and expresses it as a ratio. This ratio is obtained for each patient and is compared with the ratio obtained in the control group.

Granulocyte and mononuclear fractions were obtained from the peripheral blood of the three 5q- syndrome patients that define the 5q- syndrome critical region, and normal controls. The DNA was digested with the restriction enzyme *Eco*RI, size fractionated through 1% agarose gels and Southern blotted. Two probes were simultaneously hybridised to the filters; the insert from the I.M.A.G.E. cDNA clone and a 1.9 kb genomic *Eco*RI-*Sst*I fragment from the *renin* gene. The *renin* gene is localised to chromosome 1, uninvolved in the 5q- syndrome, and thus acts

as an internal hybridisation control. Following autoradiography, the ratio of the two signals in the patients was compared with the ratio of the two signals in the normal controls. Gene dosage experiments were carried out on two separate occasions.

3.2.5 Northern analysis

One of the I.M.A.G.E. cDNA clone inserts was hybridised to Multiple Tissue Northern (MTN) blots (Clontech) to determine the tissue expression pattern and transcript size of the cDNA. The blots contained 2µg Poly-$(A)^+$ RNA from a variety of human tissues (Table 3.1). Probes were labelled with ^{32}P-dCTP by random priming, as previously described (**Chapter 2 section 2.6**). Filters were prehybridised for 30 minutes and hybridised for 1 hour at 68°C using Expresshyb (Clontech), containing salmon sperm DNA, as previously described (**Chapter 2 section 2.7**). Filters were first washed in 2XSSC/0.5% SDS for 30 minutes at room temperature with continuous agitation and three changes of wash solution; and then in 0.1XSSC/0.1% SDS for up to 40 minutes at 50°C with continuous agitation and one change of wash solution. Autoradiography was carried out, as previously described (**Chapter 2 section 2.10**).

Table 3.1 MTN blots used in Northern analysis containing a variety of human tissues

MTN blot	Human tissues included
1	heart, brain, placenta, lung, liver, skeletal muscle, kidney, pancreas
2	spleen, lymph node, thymus, peripheral blood leukocytes, bone marrow, foetal liver

3.2.6 Southern analysis

Granulocyte DNA fractions were obtained from six 5q- syndrome patients and healthy individuals. The DNA was digested with restriction enzymes *Eco*RI, *Pst*I, *Hind*III, *Pvu*II, *and Eco*RV; size fractionated through a 1% agarose gel and Southern blotted. Two Southern blot filters were prepared and hybridised with the I.M.A.G.E. cDNA clone insert used in the gene dosage and Northern analysis, to screen for gene rearrangements. Southern analysis was carried out on two separate occasions.

3.2.7 Direct sequencing

I.M.A.G.E. cDNA clones were sequenced as either single-stranded or double-stranded templates by the dideoxy chain termination method (Sanger *et al.*, 1977), as previously described (**Chapter 2 section 2.12.1**). Clones were sequenced using the Cy5 Autoread sequencing kit (Amersham Pharmacia Biotech), which incorporated a single fluorescent label, Cy5 Amidite. Reactions were loaded onto the Alf*express* automated sequencer (Amersham Pharmacia Biotech). Data was generated for analysis using Alfwin Sequence Analyser 2.10 software (Amersham Pharmacia Biotech). Each I.M.A.G.E. cDNA clone was sequenced in full, and then subjected to a GenBank search for homology with known genes and overlapping clones to generate the full-length cDNA.

3.2.8 Overlapping cDNA clones

Sequence data from I.M.A.G.E. cDNA clones was subjected to a homology search against the EST database db(EST) to obtain overlapping cDNA clones to generate the full-length cDNA. The criteria for overlapping clones was to; have a 100% match over a ≥100bp region, and to possess an insert size large enough to extend the cDNA at either the 5' or 3' end. In addition, UniGene was accessed as it is

updated weekly with new EST sequences and bimonthly with new characterised sequences. Overlapping cDNA clones were obtained as before (**3.2.2**) and sequenced. If the overlapping clone matched the criteria, its sequence was added to the original data and the 'new' sequence submitted to db(EST) as before.

3.2.9 cDNA library screening

If no overlapping clones were identified from db(EST) or UniGene, the cDNA clone insert was screened against I.M.A.G.E. cDNA libraries from the collaboration with the Resource Centre of the German Human Genome Project at the Max-Planck-Institute for molecular genetics (RZPD (http://www.rzpd.de)). The probe was hybridised to gridded I.M.A.G.E. cDNA library filters containing a number of tissues including; liver, spleen, whole brain, skin, eye, ovary, lung, tonsil, melanocyte, pregnant uterus, heart, colon, prostate, kidney, thyroid, pancreas, and adrenal gland. Positive clones were obtained and sequenced, and the 'new' sequence submitted to db(EST) as before.

3.2.10 RACE PCR

The technology of RACE (Rapid Amplification of cDNA Ends) PCR, as described in **Chapter 2 section 2.13** was used to generate 'new' sequence when no overlapping clones were identified from screening db(EST) and cDNA libraries. RACE PCR is used to amplify the 5' and/or 3' ends of cDNAs to clone the full-length cDNA without constructing or screening a cDNA library. RACE PCR was performed using Marathon-Ready™ cDNAs (Clontech UK Ltd.). Marathon-Ready™ cDNAs are premade libraries of adaptor-ligated double-stranded cDNA ready for use as templates (Chenchik *et al.*, 1996). The libraries chosen were tissue-specific to the gene of interest.

1. The Gene-Specific Primers (GSPs) were designed only from the 5' end of the cDNA as the 3' Poly-A^+ tail had been obtained. Details of the primers, including their annealing temperatures and choice of thermal profile are shown in Table 3.2.
2. The Marathon-Ready™ cDNA templates used in the 25µl RACE PCR reaction included; human foetal liver, lung, bone marrow, pituitary gland, small intestine, foetal skeletal muscle, testis, foetal brain, foetal spleen, hypothalamus, and foetal thymus.
3. The RACE PCR products were subcloned and prepared for sequencing as previously described (**Chapter 2 section 2.14**).

3.2.11 Localisation to the YAC contig

The cDNA was sublocalised by PCR screening to the YAC contig encompassing the critical region of the 5q- syndrome (Kostrzewa *et al.*, 1998) as previously described (**Chapter 2 section 2.15**).

A PCR primer pair; Forward 1 (5'-GCTGGCACAAATGAAATGGG-3'), and Reverse 1 (5'TTGAACATGGTGTCAGTCCC-3') was designed from the cDNA sequence. PCR was performed under the following conditions: 95°C for 30 seconds, 60°C for 30 seconds, and 72°C for 1 minute. A product size of 140bp was expected.

Table 3.2 3' RACE PCR primer conditions for *C5orf4*

Primer name	Gene-specific primer sequence 5'-3'	Annealing temperature	Touchdown/three-step PCR
R1	AGGGAAACCTTTTGTTTGCG	65°C	Three-step
R2	AAAGTTTGGTCTGCACCTCG	64°C	Three-step
R3	TTTTGGTTTCCACATAAGGC	61°C	Three-step
R4	AGAAGGAGCTCTCGCTGGAGCC	73°C	Touchdown
R5	GCCTGAGTCGGGTATGTGAAAGC	70°C	Touchdown
R6	ATTTGCCAGGGTGTCTTGGCTCTG	73°C	Touchdown
R7	GCTAACTCCTGATCCCATTGAGG	69°C	Three-step
R8	CTCCAGTCTGGGGAGCTGTG	69°C	Three-step
R9	AAGGAGTTATCTGGGGCTGAACC	70°C	Touchdown
R10	AAGAAAGCCTCATCCTTGGCCCG	73°C	Touchdown
R11	GGGAAGCCGACCTTCCCTCTAC	70°C	Touchdown
R12	CCTGGTGGTGCCATCATTAGGGC	74°C	Touchdown
R13	CATCAAGGGCAGTAGTTTGAAG	63°C	Three-step
R14	CCCTGCTTTTCCGGGAAG	67°C	Three-step
R15	CTGCTTTTCCGGGAAGCCGACC	75°C	Touchdown

3.2.12 Database analysis using the Genetics Computer Group (GCG) software package

The GCG® (Genetics Computer Group) was founded in 1982 at the Department of Genetics at the University of Wisconsin-Madison. The Wisconsin Package Version 10.1 is an integrated package of over 130 programs that allows the manipulation and analysis of nucleic acid and protein sequences.

3.2.12.1 FastA

FastA uses the method of Pearson and Lipman, (1988) to search for similarities between a query sequence and a group of sequences of the same type (nucleic acid or protein). The 3051bp sequence of the novel cDNA was submitted to a FastA search in the GenEMBL group of sequence databases.

3.2.12.2 BLAST

BLAST (Basic Local Alignment Search Tool) uses the method of Altschul *et al.*, (1990) to search one or more nucleic acid or protein databases for sequences similar to one or more query sequences of any type. The novel cDNA was submitted to a BLAST search in the SWISS-PROT protein sequence database.

3.2.12.3 Frames

Frames displays open reading frames for the six translation frames of a DNA sequence. The open reading frame of the novel cDNA was determined using the X-Windows graphics display software run on 'eXodus' for the Apple Macintosh.

3.2.12.4 Translate

Translate translates nucleotide sequences into peptide sequences. A translation was carried out to determine the coding region of the cDNA in the correct open reading frame.

3.2.12.5 Motifs

Motifs looks for sequence motifs by searching through proteins for the patterns defined in the PROSITE Dictionary of Protein Sites and Patterns. A Motif search was carried out on the translated region of the novel cDNA.

3.2.12.6 GenBank submission

GenBank (http://www.ncbi.nlm.nih.gov/GenBank/index.html) is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences. There are approximately 15,850,000,000 bases in 14,976,000 sequence records as of December 2001. Many journals require submission of sequence information to a database prior to publication so that an accession number may appear in the paper. NCBI has a www form called BankIt (http://www.ncbi.nlm.nih.gov/BankIt/index.html), for convenient and quick submission of sequence data.

3.2.13 Mutation analysis of *C5orf4*

The method of choice for mutation analysis on the translated sequence of novel gene *C5orf4* was cycle sequencing using the Thermo Sequenase™ Cy™5 Dye Terminator Kit (Amersham Pharmacia Biotech) on the Alf*express* automated sequencer (Amersham Pharmacia Biotech). This was due to the small coding region of the gene (432bp) which could be cycle sequenced in one fragment only.

As *C5orf4* was novel, the genomic structure of the gene was unknown, therefore a Reverse Transcriptase (RT-PCR) was performed to generate the cDNA template. The cDNA template was then combined with a gene-specific unlabelled primer, Thermo Sequenase DNA polymerase, Cy5 labelled ddNTPs, and dNTPs. Thermo Sequenase gives an even incorporation of the nucleotides and generates very uniform signal peaks.

3.2.13.1 Samples

Ten patients with the 5q- syndrome were included in the study. Granulocyte and mononuclear cells were separated from 40mls of peripheral blood by ficoll gradient centrifugation (Boyum, 1984). The granulocytes showed a high level of purity (≥95%). Mononuclear cells showed a purity of ≥90%. Total RNA was obtained from the granulocyte fractions with the Total RNA Isolation Kit (AMS Biotechnology (Europe) Ltd. This method is based on the disruption of cells in guanidinium thiocynate/cationic detergent solutions, followed by the organic extraction and alcohol precipitation of the RNA. Granulocyte total RNA fractions from the peripheral blood of healthy individuals were used as controls.

3.2.13.2 Reverse Transcriptase PCR (RT-PCR)

RT-PCR was carried out using the R*everse*--iT™ One-step PCR kit (ABgene) as previously described in **Chapter 2 section 2.16**.

An RT-PCR primer pair; Forward 3 (5'-TGCAGATGGTCAGGATGG-3'), and Reverse 3 (5'-TCTGACCAGCCTCCAGTC-3') was designed flanking the coding region of *C5orf4*. The primers were approximately 50% GC and at least 19 bases in length. Additionally, the primers contained either a G or C residue as the last 3'-base, and did not have any regions that could self-anneal or form "hair pin" loops. The primers were dissolved in RNase-free water to a concentration of 100pm/µl.

The RT-PCR primers annealed to the template at 63°C. A product size of 550bp was expected.

3.2.13.3 Cycle sequencing

Cycle sequencing reactions were carried out as described in **Chapter 2 section 2.18.**

Cycle sequencing primers; Forward 4 (5'-AAATGTGTGAGGCTGGCAC-3'), and Reverse 4 (5'-TCTGGGGAGCTGTGTTTTC-3') were designed internal to the RT-PCR primers, but still flanking the coding region of novel gene *C5orf4*. Both primers annealed to the template at 61°C in the sequencing reaction.

3.2.14 Database analysis using the Genetics Computer Group (GCG) software package

Sequence data from each patient and control from novel gene *C5orf4*, was compared with the sequence data submitted to GenBank using BestFit analysis.

3.2.14.1 BestFit analysis

BestFit makes an optimal alignment of the best segment of similarity between two sequences. Optimal alignments are found by inserting gaps to maximise the number of matches using the local homology algorithm of Smith and Waterman.

3.3 Results

3.3.1 ESTs

The Human Gene Map of chromosome 5 between DNA markers D5S410 and D5S487 at NCBI contained an unidentified transcript, Hs.10235 represented by 21 ESTs. Two of these ESTs (GenBank Accession Numbers: AA029816 and N73983) were selected for analysis.

3.3.2 I.M.A.G.E. cDNA clones

I.M.A.G.E. cDNA clones 469867 and 296617 from GenBank Accession Numbers AA029816 and N73983 respectively were obtained as stabs in agar from the HGMP-RC. cDNA clone 469867 was identified from the Soares pregnant uterus NbHPU *Homo sapiens* cDNA library while cDNA clone 296617 was identified from the Soares foetal liver spleen 1NFLS cDNA library.

3.3.3 Gene dosage analysis

Gene dosage analysis with cDNA clone 469867 and the *renin* gene showed both probes hybridised to a single fragment. An approximate 50% reduction in the dosage of clone 469867 in the granulocyte patient DNA, compared with normal controls confirmed the deletion of one allele, and therefore, co-localisation to 5q.

3.3.4 Northern analysis

Northern analysis showed I.M.A.G.E. cDNA clone 469867 to possess a single transcript of 3.0kb and to be ubiquitously expressed with a high level of expression in foetal liver, see Figure 3.1.

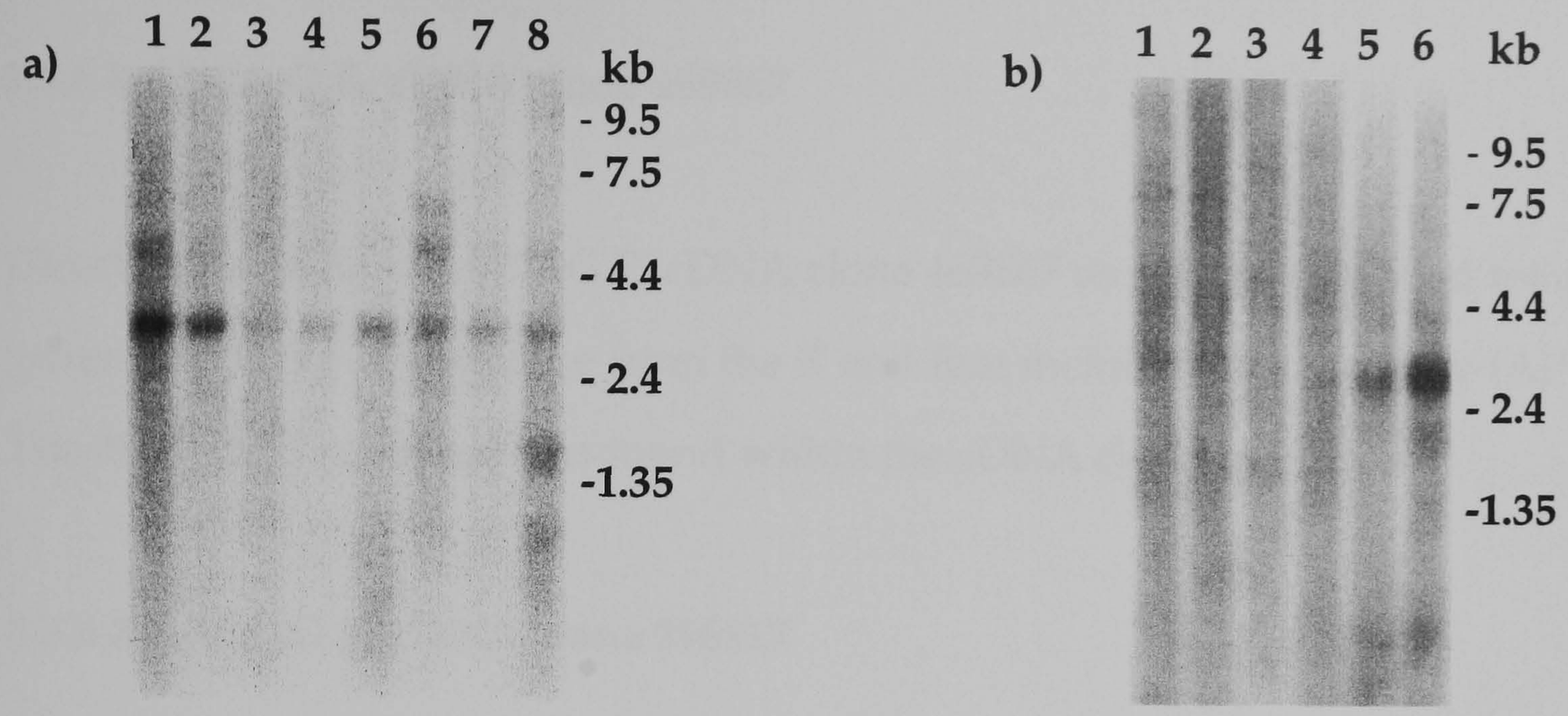


Figure 3.1

Representative Northern blot analysis of I.M.A.G.E. cDNA clone 469867. MTN blot (a) included 2µg of Poly-(A)$^+$ RNA from; heart (1), brain (2), placenta (3), lung (4), liver (5), skeletal muscle (6), kidney (7), and pancreas (8). MTN blot (b) included 2µg of Poly-(A)$^+$ RNA from; spleen (1), lymph node (2), thymus (3), peripheral blood leukocytes (4), bone marrow (5), and foetal liver (6). Sizes of RNA marker bands (kb) are indicated approximately.

3.3.5 Southern analysis

No rearrangements were observed in the granulocyte fractions of six patients with the 5q- syndrome digested with restriction enzymes *Eco*RI, *Pst*I, *Hind*III, *Pvu*II, and *Eco*RV, following hybridisation with cDNA clone 469867.

3.3.6 Direct sequencing

3.3.6.1 I.M.A.G.E. cDNA clone 469867

Direct sequencing of I.M.A.G.E. cDNA clone 469867 as a single-stranded template generated 332bp of sequence from the 3' end that included an 18bp Poly-$(A)^+$ tail. The 280bp EST sequence was found within the cDNA clone as expected.

3.3.6.2 I.M.A.G.E. cDNA clone 296617

Direct sequencing of I.M.A.G.E. cDNA clone 296617 as a double-stranded template generated 1500bp of sequence from the 3' end that included an 18bp Poly-$(A)^+$ tail. The 261bp EST sequence was not found within the cDNA as expected, suggesting the clone was incorrect. Clone 296617 was discarded from this point.

3.3.7 Overlapping cDNA clones

A db(EST) search using the sequence of 469867 identified two overlapping cDNA clones (I.M.A.G.E. No.: 982453, GenBank Accession No.: AA523234; I.M.A.G.E. No.: 935769, GenBank Accession No.: AA523935) from the db(EST) homology search. They were derived from the metastatic prostate bone lesion NCI_CGAP_Pr12 and the colon NCI_CGAP_Co3 cDNA libraries respectively. These two clones had the potential to add to the sequence of clone 469867 and contribute to the construction of the full-length cDNA.

3.3.7.1 I.M.A.G.E. cDNA clone 982453

The db(EST) search with cDNA 469867 revealed a 100% match with clone 982453 at the 3' end immediately preceding the Poly-$(A)^+$ tail. Direct sequencing of clone 982453 as a double-stranded template generated 221bp of sequence at the 3' end within the 332bp sequence of clone 469867. Therefore, no further sequence could be added to cDNA 469867.

3.3.7.2 I.M.A.G.E. cDNA clone 209846

The db(EST) search on overlapping clone 982453 (**3.3.7.1**) identified one overlapping clone, recently deposited, (I.M.A.G.E. No.: 209846, GenBank Accession No.: H67084) from the Soares foetal liver spleen 1NFLS cDNA library to be matching 100% with cDNA 469867 at the 3' end, immediately preceding the Poly-$(A)^+$ tail. Information data on clone 209846 showed it to possess a large insert size of 1391bp that would extend cDNA clone 469867 a further 1059bp. Direct sequencing of clone 209846 as a single-stranded template generated 1259bp of sequence, and the EST data was found within the sequence. The sequence overlapped with 469867 with 100% homology over 332bp. Thus, clone 209846 added a further 927bp to the cDNA. A second db(EST) search identified a further four overlapping clones (I.M.A.G.E. No.s.: 280058, 292429, 241086, and 203279) which would potentially extend the 1259bp cDNA, see Table 3.3.

Table 3.3 Overlapping I.M.A.G.E. cDNA clones identified from db(EST) homology searches from I.M.A.G.E. cDNA clon 469867

I.M.A.G.E. cDNA clone	GenBank Accession No.	cDNA library source	Size of insert (bp)	EST sequence present?	Overlap with 469867 (bp)
982453	AA523234	Metastatic prostate bone lesion NCI_CGAP_Pr12	221bp	Yes	221bp
935769	AA523935	Colon NCI_CGAP_Co3	449bp	No	None
209846	H67084	Soares foetal liver spleen 1NFLS	1391bp	Yes	332bp
251760	H97862	Soares melanocyte 2NbHM	2525bp	No	None
280058	N56931	Soares multiple sclerosis 2NbHMSP	1300bp * #	Yes	250bp
292429	N68417	Soares foetal liver spleen 1NFLS	1000bp *	Not sequenced	Not sequenced
241806	H93077	Soares foetal liver spleen 1NFLS	1200bp *	Not sequenced	Not sequenced
203279	H54756	Soares foetal liver spleen 1NFLS	1200bp *	Not sequenced	Not sequenced
462172	AA705416	Soares foetal liver spleen 1NFLS_S1	Not known	No	None
275607	R93303	Soares foetal liver spleen 1NFLS	1046bp	No	None
435297	AA699911	Soares foetal liver spleen 1NFLS_S1	1428bp	Yes	52bp

* Approximate insert sizes calculated from a 1% agarose gel
\# 1164bp on sequencing

3.3.7.3 I.M.A.G.E. cDNA clone 280058

The information on clones 280058, 292429, 241086, and 203279 did not include insert sizes, therefore inserts were cut out using the appropriate restriction enzymes. Clone 280058 possessed the largest insert with a size of 1300bp, and was thus chosen for further analysis. Direct sequencing of cDNA 280058 as a single-stranded template showed it to overlap 100% at the 5' end from nucleotide 82, with the cDNA sequence. However, the clone was not large enough to add any further sequence data to the cDNA, and was thus discarded at this point.

3.3.7.4 False positive I.M.A.G.E. cDNA clones 935769 and 251760

The db(EST) search with cDNA 469867 revealed clone 935769 to be matching 100% with cDNA 469867 at the 3' end immediately preceding the Poly-$(A)^+$ tail. Direct sequencing of clone 935769 as a single-stranded template showed it did not contain its EST sequence, and thus did not overlap with cDNA 469867, confirming it was the wrong clone. The cDNA was discarded from this point. The db(EST) homology search on the 'new' 1259bp sequence (469867 + 209846) revealed one overlapping clone that would potentially complete the full coding sequence (cds). The search showed clone 251760 to be matching 100% over its 567bp EST sequence at the 3' end immediately preceding the Poly-$(A)^+$ tail. Direct sequencing of the double-stranded template showed it did not contain its EST sequence, and thus did not overlap with the cDNA. Clone 251760 was discarded from this point.

3.3.8 cDNA library screening

The collaboration with the Resource Centre of the German Human Genome Project came into effect when a db(EST) homology search failed to identify overlapping clones. Thirty-four potential positive clones were identified with probe 209846 (**3.3.7.2**), see Table 3.4. Direct sequencing of the clones identified 15

positive clones out of 28 analysed, see Table 3.4. Eight clones were not screened because their insert sizes were less than 500bp. None of the positive clones added any further sequence to the cDNA.

3.3.9 RACE PCR

RACE PCR was used to extend the cDNA after db(EST) homology searches, UniGene, and cDNA library screening failed to produce any further clones.

Firstly, 3 GSPs were designed (R3-R1) approximately 200bp from the 5' end of the sequence. A 294bp RACE PCR product was generated from the human foetal brain cDNA library. The product was purified, subcloned, and the plasmid extracted. Direct sequencing of the double-stranded template showed it to have a 100% match with the cDNA, in a 129bp overlap, and extend the sequence by 165bp. No overlapping clones were identified from db(EST) and UniGene homology searches. Therefore, RACE PCR primers were designed from the 'new' sequence. The second RACE PCR product was generated from the human foetal brain cDNA library also. It had a 100% match with the cDNA, overlapped by 76bp, and extended the sequence by 164bp. The transcript sequence generated was now 1588bp.

Table 3.4 **Clones identified from screening the I.M.A.G.E. cDNA clone collection library with probe 209846**

Clone	RZPD Clone ID	Approx insert size	Clone Status	Overlap with cDNA
A1	M0978Q2	900bp	Positive	Yes (900bp)
A2	K1479Q2	924bp	Positive	Yes (924bp)
A3	J17114Q2	1000bp	Positive	Yes (1000bp)
A4	B19370Q2	926bp	Positive	Yes (926bp)
A5	P11389Q2	997bp	False positive	No
A6	E01393Q2	759bp	Positive	Yes (759bp)
B1	M21395Q2	2700bp	False positive	No
B2	E20396Q2	700bp	Positive	Yes (700bp)
B3	M14409Q2	?	False positive	Yes
B4	M15409Q2	1259bp	Positive	Yes (1259bp)
B5	A03414Q2	CLONE	NOT	SCREENED
B6	B08414Q2	735bp	False positive	No
C1	B09414Q2	732bp	False positive	No
C2	C03414Q2	1238bp	NOT	SCREENED
C3	C08414Q2	723bp	NOT	SCREENED
C4	D03414Q2	962bp	NOT	SCREENED
C5	J01414Q2	?	False positive	No
C6	J03414Q2	?	False positive	No
D1	G21463Q2	1000bp	Positive	Yes (1000bp)
D2	G09516Q2	1200bp	RECOMBINANT	CLONE
D3	G14517Q2	1706bp	False positive	No
D4	A07520Q2	1000bp	Positive	Yes (1000bp)
D5	F18528Q2	CLONE	NOT	SCREENED
D6	G15528Q2	2329bp	False positive	No
D7	L19536Q2	1489bp	False positive	No
D8	F15608Q2	1033bp	Positive	Yes (1033bp)
D9	J13654Q2	CLONE	NOT	SCREENED
D10	B17656Q2	CLONE	NOT	SCREENED
D11	B04536Q2	900bp	Positive	Yes (900bp)
D12	L24657Q2	1200bp	Positive	Yes (1200bp)
D13	M02662Q2	1000bp	Positive	Yes (1000bp)
D14	F14608Q2	1705BP	False positive	No
D15	N12651Q2	CLONE	NOT	SCREENED
D16	G03535Q2	826bp	Positive	Yes (826bp)

3.3.10 I.M.A.G.E. cDNA clone 435297

The db(EST) homology search on the 1588bp sequence identified an overlapping clone (I.M.A.G.E. No.: 435297, GenBank Accession No.: AA699911) derived from the Soares foetal liver spleen 1NFLS_S1 cDNA library. Restriction enzyme digestion showed the insert size to be approximately 1.4kb. This clone would complete the full cds. However, the 100% overlap of clone 435297 with the cDNA was only 52bp. The ideal overlap is between 80-100bp to confirm the clone is part of the cDNA.

Southern and Northern data to confirm the clone was part of the cDNA was inconclusive. RACE PCR primers (R10-R12) were designed. The 144bp RACE product showed 100% homology with cDNA clone 435297, confirming the clone was part of the cds. Direct sequencing of clone 435297 as a double-stranded template with internal primers generated a 1428bp fragment that completed the full-length cDNA. I.M.A.G.E. cDNA clones 1654025 (GenBank Accession No.: AI084848), and 121027 (GenBank Accession No.: T96312), both from the Soares foetal liver spleen 1NFLS cDNA library and possessing insert sizes of 653bp and 1058bp respectively were directly sequenced as double-stranded templates to confirm the sequence of clone 435297 and erase any sequencing ambiguities.

3.3.11 Localisation to the YAC contig

Novel gene *C5orf4* was sublocalised by PCR screening to YAC 914A10 from the YAC contig spanning the critical region of the 5q- syndrome (Kostrzewa *et al.*, 1998), see Figure 3.2

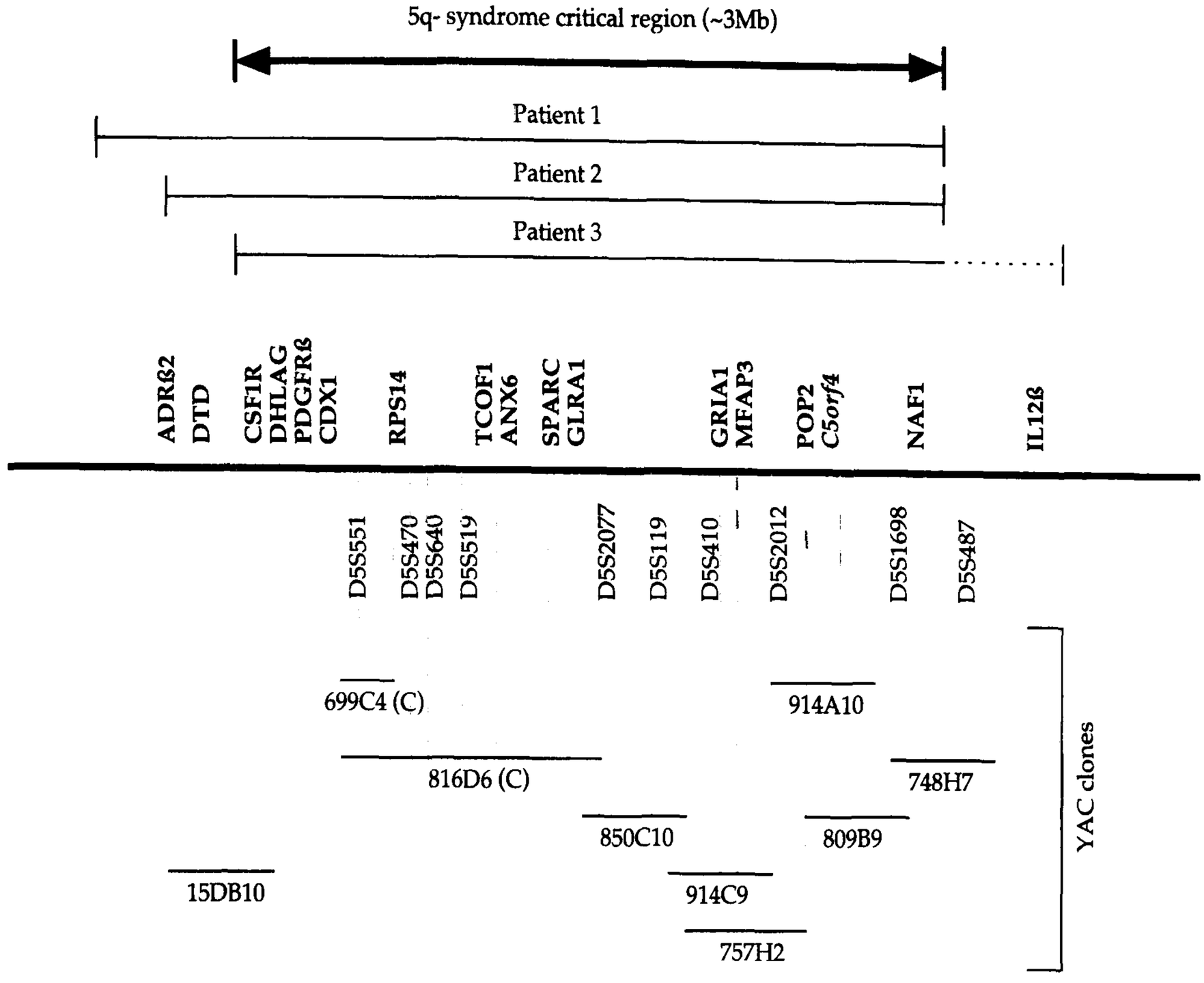


Figure 3.2

Transcription map of the critical region of the 5q- syndrome. The map shows known genes (boldface) and novel gene C5orf4 (italics) cloned in this study. The three 5q- syndrome patients that define the critical region of gene loss are also shown. C5orf4 is shown to map to YAC 914A10 from the YAC contig encompassing the critical region of the 5q- syndrome (Kostrzewa *et al.*, 1998). C, denotes chimaeric YAC.

3.3.12 Database analysis using the Genetics Computer Group (GCG) software package

GCG was accessed via telnet at the National Centre for Supercomputing Applications (NCSA), and locally at the Oxford University Bioinformatics centre.

3.3.12.1 FastA analysis

The FastA nucleotide homology search utilised all databases (GenEMBL). The 3051bp cDNA showed no known homology with any gene, identifying it as completely novel.

3.3.12.2 BlastX analysis

The BlastX protein homology search from the SWISS-PROT database showed the 3051bp cDNA to have no known protein homology with any known gene.

3.3.12.3 Frames analysis

Frames was carried out on the novel cDNA to determine the open-reading frame (ORF) for the sequence. The figure indicated it to be in ORF +1 as this frame showed the longest stretch of translated sequence.

3.3.12.4 Translation of the novel cDNA

Confirmation of the Frames prediction was achieved by translating the cDNA. Translation was carried out on all three ORFs starting from nucleotides 1, 2, and 3 respectively to determine the 5' and 3' untranslated regions (UTRs), and the coding region of the novel cDNA. Translation in ORF 1 revealed a Methionine (M) at nucleotide position 993 (amino acid position 1) and a Stop codon (*) at nucleotide position 1425 (amino acid position 475). This indicated a putative coding region of 432 nucleotides (144 amino acids) in Frame 1, see Figure 3.3.

```
   1  aaggaagtttgcctctggtaccaccttgggactaagtgtatacccaactaagtcccttca   60
  61  tctccctggaccctgccttcaatcatccctcctcttgcaggtctccaacatgctaccggt  120
 121  gatagtgggcccattagtaatgggctcccacttgtcctccatcaccatgtggttttcctt  180
 181  ggccctcatcatcaccaccatctcccactgtggctaccaccttcccttcctgccttcgcc  240
 241  tgaattccacgactaccaccatctcaagtaaggaccttctccccacaatgggtccctagg  300
 301  caacaaccatcaccctccctccttaagcttctgatagcaggcactggtatctcagtttac  360
 361  atgtgagtgctagttggggaagaggtctgatggctgggaaaataaccacactgggatgac  420
 421  cttactcttttccttctgaaattgggtcctcctccaactgggggttttaaacctatattt  480
 481  ggactctgggccctttgaataagctctggacattctctttagaaaagtacatatatgcat  540
 541  aaaactcctgcaggcaatcacaagtaagtcatgtatctctgaaagccccagacttttca  600
 601  agtgtagtctacaaacctcctataccagaatcacctgagatgcttgtttaaaatataggc  660
 661  tcttgaattccactccaaatatgctgaactaaggtatctgggggagcagggcctggaaac  720
 721  ctgcattttagttaagtgttccaggggattctgatgctggtgatctcatacctgaagaag  780
 781  cctagttctgctagagattctcagactgtgggcatcaatcacatagcaagcttttcaaag  840
 841  gaggactccagatacggcaaattgcaggtttggagtggcccccgggagatgcattggtag  900
 901  caactccccagcttattctgatgcagatggtcaggatggtggtgacaggtgtggggtagg  960

 961  agggctgagaaatgtgtgaggctggcacaaatgaaatgggatctgcacggaagcctccag 1020
                                      M  K  W  D  L  H  G  S  L  Q
1021  catgggcactgcagtgccgtcctctcttcttctaggttcaaccagtgctatggggtgctgg 1080
      H  G  H  C  S  A  V  L  S  L  L  G  S  T  S  A  M  G  C  W
1081  gtgtgctggaccacctccatgggactgacaccatgttcaagcagaccaaggcctacgaga 1140
      V  C  W  T  T  S  M  G  L  T  P  C  S  S  R  P  R  P  T  R
1141  gacatgtcctcctgctgggcttcaccccgctctctgagagcatcccagactccccaaaga 1200
      D  M  S  S  C  W  A  S  P  R  S  L  R  A  S  Q  T  P  Q  R
1201  ggatggagtgagagacagcctaagtgtcatcctggctgtccctcagccatgggatgcaga 1260
      G  W  S  E  R  Q  P  K  C  H  P  G  C  P  S  A  M  G  C  R
1261  cacggcttcctgattgcacctaacaatttgcctccttcggccacacgccctaatgatggc 1320
      H  G  F  L  I  A  P  N  N  L  P  P  S  A  T  R  P  N  D  G
1321  accaccagggtagagggaaggtcggcttcccggaaaagcagggccaaggatgaggctttc 1380
      T  T  R  V  E  G  R  S  A  S  R  K  S  R  A  K  D  E  A  F
1381  ttcaaactactgcccttgatgtccctcaatgggatcaggagttagcttaagaaaaaggaa 1440
      F  K  L  L  P  L  M  S  L  N  G  I  R  S  *

1441  aacacagctccccagactggaggctggtcagagggaggagacccctggtcctctgctgtg 1500
1501  gaaggagaggggttcagccccagataactcctttgtggcctgggcaggatgcagagaatg 1560

1561  acaaggctgaaaggaggggactggaggctgcctggctccagcgagagctccttctggga 1620
1621  ccagagggtgggacggccaggtatcactttgccccttcctgccccaaaaggctttcacat 1680
1681  acccgactcaggccagagccaagacaccctggcaaatcattataggtctcaattcatgac 1740
1741  ataccagatgccagtcgccacttcacccaacacacacacaaacatacacacaccaaacac 1800
1801  tctgagtgagtggtaaaggccccgtttttactctggcccaccgcaaacaaaaggtttccc 1860
1861  tctgtgggggagaaaaagaaatccaggagctcctccctggattaaaaccaacgaggtgca 1920
1921  gaccaaactttaacacctttagccttatgtggaaaccaaaaaccaactgctgggaaactg 1980
1981  tgaaaagccctttacccacaaggggaggggtcaaagttgctgccctttggggacacccg 2040
2041  agacccctaattagcctatctgaatgaggaccaaaggttagagccctctttctccccgaa 2100
2101  gaaagagcccggaaaaacatggcagagcaaagagcaaaatcctttctcccccgaatgctt 2160
2161  taccagtttctcagcaacatttattcaagatgattttaccaggaaccttatcaaaggca 2220
2221  aaaccacagctgcttgggttgaagtccccatcctggcctcctctcagccgccagacatgg 2280
2281  ccaggaaccctgtggttcccaagaacaatttaaagatcactctttgatttgaaagaccac 2340
2341  cattatcattttactaaattcttatatatacttgtgccttttccaactttcagttctctt 2400
2401  aagaaaaacatccactgtagctttataaaataccttagtatcagctaggtccaagtctcc 2460
2461  aggcaggactatcaaaatggactccttcttcctatcagctctagaaccccagaggacttg 2520
2521  cccagtaggccagctgagcacttaccaggcagcatctccctggccttccacagtctcgcc 2580
2581  tgtctgtctctggccttggtggacctggctccctccctaggaggcttttgccctcagctt 2640
2641  gaatacagttcctggtgcatcaaagctgtaagttctcagctgctggcctgcactcatcac 2700
2701  ctcttcctacaaataaacatttggaaaaaagtccatcttcaatatgcttaaagagaaggg 2760
2761  taggagataaggagaaagaacagatacggtttttttccccctttaaggcctttagattttg 2820
2821  aggtactgtaaggggcctagaaggacaaaaggctttcattccctcttccttttggcaggc 2880
2881  aggttatcagtccttggcagaagggcccagccctatccttttctgtatcaaacaaaatcc 2940
2941  ttgaggttggtatacaagttaaggctgaaaaaaggccttaaattcccagtaaagaatgtg 3000
3001  aaagcaagcatgtaaaataaactggtcttcatgaaaaaaaaaaaaaaaaaa          3051
```

Figure 3.3 Nucleotide and predicted amino acid sequence of *C5orf4*. The cDNA contains a 144 amino acid ORF. Numbers indicate nucleotide positions. The predicted translational start codon (ATG), the predicted translational stop codon (TAG), and the putative polyadenylation signal (AATAAA) are underlined.

3.3.12.5 Motif search

A motif search was carried out on the 144 amino acid protein sequence of the cDNA. No significant motif sequences were found.

3.3.12.6 GenBank submission

The novel cDNA was assigned the name *C5orf4* based on its chromosomal localisation (C5) and being the next one in the series (orf4) by the Human Gene Nomenclature Committee, The Galton Laboratory, University College London, UK (http://www.gene.ucl.ac.uk/nomenclature/). The data was submitted using the www BankIt form at NCBI, and released on the public database on January 1, 2000.

3.3.13 Mutation analysis of *C5orf4*

No mutations were found in the 432bp coding region of novel gene *C5orf4* in 8 patients with the 5q- syndrome included in the study, see Figure 3.4.

It is therefore unlikely that novel gene *C5orf4* is the tumour suppressor gene associated with the development of the 5q- syndrome.

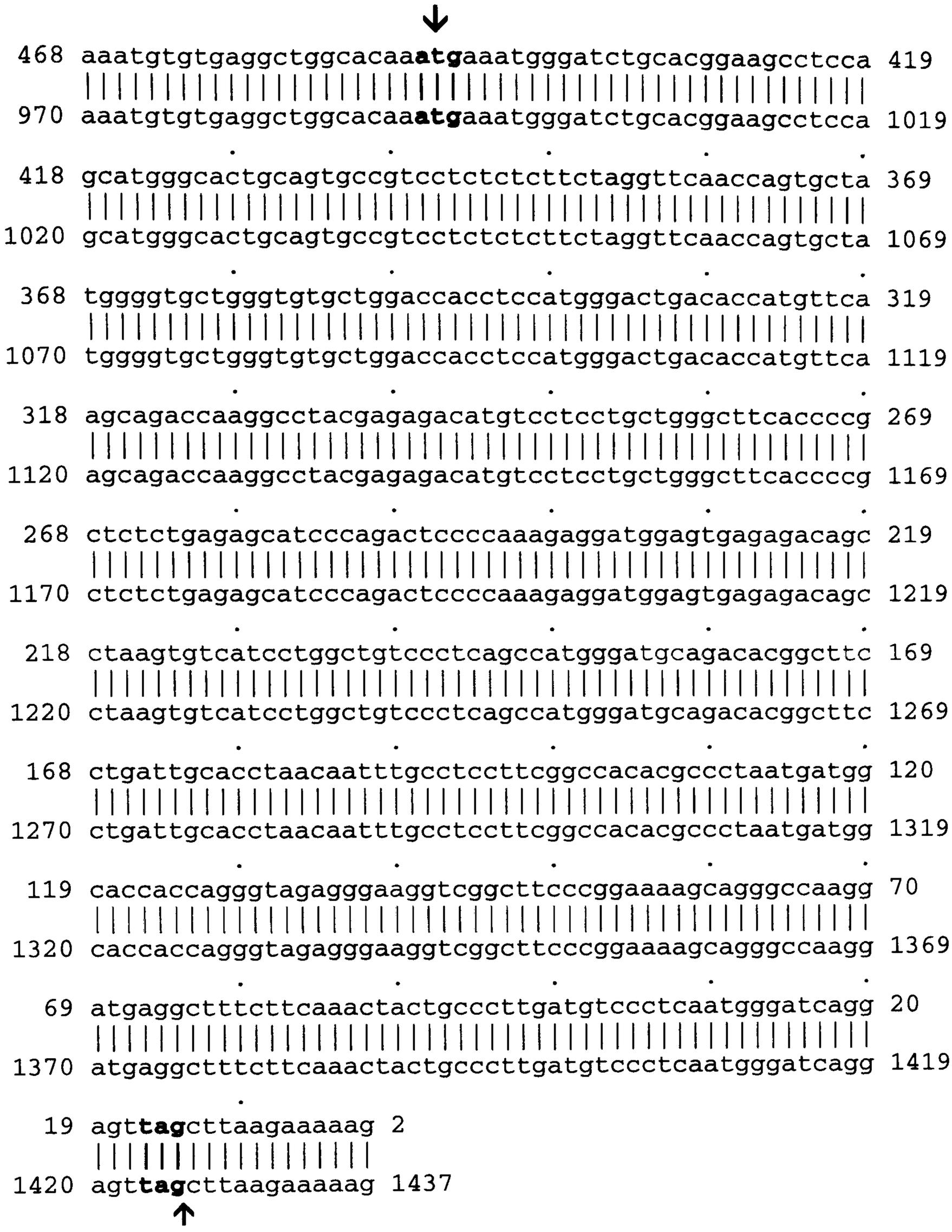


Figure 3.4

Representative mutation analysis by cycle sequencing of patients with the 5q- syndrome, and normal controls. Alignment of patient 3 (top line) and *C5orf4* (bottom line) from the 432 bp coding region of the novel gene. The arrows indicate the translational start codon (**atg**) and the translational stop codon (**tag**). Patient 3 has 100% homology with the *C5orf4* gene indicating no mutations were found.

3.4 Discussion

The EST database db(EST) was used as the resource to identify novel genes mapping to the critical region of the 5q- syndrome at 5q31.3-q33. The Human Chromosome 5 Gene Map was accessed between DNA markers D5S410 and D5S487 that span the critical region of gene loss. An unidentified transcript represented by 21 EST sequences was identified. Novel gene *C5orf4* was cloned from three cDNA clones and two RACE PCR products. The cloning of *C5orf4* has contributed to the HGP in the identification of the estimated 80-100,000 (now believed to be 30-40,000) disease genes in the human genome, and represents a candidate gene for the putative tumour suppressor gene associated with the development of the 5q- syndrome.

3.4.1 Molecular studies

Novel gene *C5orf4* was shown to map to the critical region of gene loss by gene dosage analysis, and then sublocalised by PCR screening to YAC 914A10 from the YAC contig encompassing the critical region of the 5q- syndrome (Kostrzewa *et al.*, 1998). Northern analysis showed *C5orf4* to possess a single transcript of 3.0kb and to be ubiquitously expressed with a high level of expression in foetal liver, suggesting a possible role in early haematopoiesis. The Wilms' tumour gene, *WT1*, which is essential for kidney development, and mutated in some Wilms' tumours is an example of a gene thought to play a role in early haematopoiesis. Studies have shown *WT1* to have a high level of expression in progenitor cells (namely $CD34^+$), to be downregulated during the differentiation of leukaemic cell lines, and high levels of *WT1* expression to cause cell-cycle arrest, and/or apoptosis (Pritchard-Jones and King-Underwood, 1997). This may reflect a role in the control of normal haematopoiesis, which can be altered by mutations in the gene

and form part of the pathway towards leukaemogenesis. The foetal liver expresses *WT1* at a time when it is a site of active haematopoiesis. Another example is the GATA-1 transcription factor that is expressed in early haematopoiesis. *In vitro* experiments suggest that its transcription is activated by two specific enhancers only in haematopoietic cell lines, thus involved in haematopoietic-specific regulation (Wu *et al.*, 1995).

Therefore, the localisation, expression pattern, and function hypothesis would suggest novel gene *C5orf4* to be a candidate for the putative tumour suppressor gene associated with the development of the 5q- syndrome.

3.4.2 Translation of *C5orf4*

The full coding sequence of novel gene *C5orf4* was generated by sequencing overlapping I.M.A.G.E. cDNA clones, screening cDNA libraries, and RACE PCR. Translation of *C5orf4* defined a 990bp 5' UTR; predicted a 144 amino acid protein with a Methionine START codon and a STOP codon, and a 1626bp 3' UTR including a putative polyadenylation site. No sequence motifs were found in the *C5orf4* predicted protein following a database search of the PROSITE dictionary of protein sites.

C5orf4 has a small open reading frame of 144 amino acids. However, there are other novel genes with a similar size coding region cited in the literature. One example is the novel transcript *INE1* (Inactivation Escape 1) which has been localised to chromosome interval Xp21.1-p11.23, and escapes X-inactivation. This cDNA defines a complete sequence of 941bp with a predicted protein of 52 amino acids including a Methionine START codon and a STOP codon (Esposito *et al.*, 1997). The complete sequence of novel gene *C5orf4* was submitted to GenBank and assigned the Accession No. AF159165 (Boultwood *et al.*, 2000).

3.4.3 Mutation analysis of *C5orf4*

To investigate the proposal that novel gene *C5orf4* may be mutated in the 5q- syndrome, we cycle sequenced DNA samples from eight patients with the 5q- syndrome for mutations in the coding region of *C5orf4*. No mutations were found in the coding region of the *C5orf4* gene in the eight patients included in the study.

3.5 Conclusion

The EST resource has been successfully used to identify novel gene *C5orf4* that maps to the critical region of the 5q- syndrome at 5q31.3-q33. Database searches showed *C5orf4* had no protein homology to any known eukaryotic gene, identifying it as completely novel. However, its expression in $CD34^+$ cells, and its high expression in foetal liver suggests a possible role in early haematopoiesis. *C5orf4* was therefore screened for mutations by cycle sequencing in eight patients with MDS and a 5q deletion. No mutations were observed in the coding region of novel gene *C5orf4*, suggesting it was unlikely to be the tumour suppressor gene involved in the development of the 5q- syndrome.

Since this study was completed, a new patient with MDS and the 5q- syndrome has been identified. Boultwood and Fidler *et al.*, (in press) identified a fourth patient with the 5q- syndrome and a small deletion to refine further the critical deleted region. This resulted in the narrowing of the CDR at the distal breakpoint at 5q32 to approximately 1.5Mb at 5q31.3-5q32, flanked by the DNA marker D5S413 and the glycine receptor (*GLRA1*) gene. The distal breakpoint of the new critical region of the 5q- syndrome now excludes novel gene *C5orf4*, and thus confirms that *C5orf4* is unlikely to be the tumour suppressor gene associated with the development of the 5q- syndrome.

Chapter 4

Identification, localisation and analysis of novel cDNAs mapping to the critical region of the 5q- syndrome

4.1 Introduction

4.1.1 Identifying novel genes implicated in disease

4.1.2 Techniques for isolating novel coding sequences

4.1.2.1 Exon trapping

4.1.2.2 Direct selection

4.1.2.3 YAC hybridisation

4.1.3 Database searching and ESTs

4.1.4 Isolating novel genes mapping to the critical region of the 5q- syndrome

4.2 Materials and Methods

4.2.1 EST selection

4.2.2 I.M.A.G.E. cDNA clones

4.2.3 Samples

4.2.4 Gene dosage analysis

4.2.5 Northern analysis

4.2.6 Southern analysis

4.2.7 Direct sequencing

4.2.8 Overlapping cDNA clones

4.2.9 cDNA library screening

4.2.10 Dot blot analysis

4.2.11 RACE PCR

4.2.12 Localisation to the YAC contig

4.2.13 Expression analysis

4.2.14 Database analysis using the Genetics Computer Group (GCG) software package

4.3 Results
4.3.1 EST identification
4.3.2 Gene dosage analysis
4.3.3 Northern analysis
4.3.4 cDNA library filter hybridisation
4.3.5 Dot blot analysis
4.3.6 Direct sequencing
4.3.7 Southern analysis
4.3.8 Overlapping cDNA clones
4.3.9 cDNA library screening
4.3.10 RACE PCR
4.3.11 Localisation to the YAC contig
4.3.12 Expression analysis
4.3.13 Database analysis using the Genetics Computer Group (GCG) software package
4.3.14 Novel cDNA AF010242
4.3.15 Novel cDNA AF156165
4.3.16 Novel cDNA H82831
4.3.17 Summary
4.4 Discussion
4.4.1 Mapping of chromosome 5-specific ESTs to the YAC contig
4.4.1.1 Novel cDNA AF010242
4.4.2 Novel gene identification from the Human Chromosome 5 GeneMap'98
4.4.2.1 Novel gene AF156165
4.4.3 Novel gene identification from the Human Chromosome 5 GeneMap'99
4.4.3.1 Novel cDNAs AA040631, R76720, and H82404
4.4.4 Redundancy in the EST database
4.4.5 Human Genome Project progress
4.5 Conclusion

4.1 Introduction

4.1.1 Identifying novel genes implicated in disease

The identification of novel coding sequences in the human genome has been important in the discovery of genes implicated in cancer and disease. Current strategies including linkage analysis, cytogenetic analysis (cloning a chromosome translocation), and LOH have been used to identify some of these novel genes. An example of a novel gene identified by linkage analysis is the *APC* gene. Familial adenomatous polyposis (FAP) is an autosomal dominant condition characterised by diffuse intestinal polyposis, specific gene mutation, and predisposition for developing colon cancer. DNA from sixty-one unrelated patients with FAP was examined for mutations in three genes (*DP1*, *SRP19*, and *DP2.5*) located within a 100kb region deleted in two of the patients (Groden *et al.*, 1991). Four mutations were found by single-strand confirmation polymorphism (SSCP) analysis in the exons of the *DP2.5* gene. Analysis of the DNA from the parents of one of these patients showed that this 2bp deletion is a new mutation. Moreover, the mutation was transmitted to two of his children. These data have established that *DP2.5* is the *APC* gene.

Familial cases of MDS are rare. Mandla *et al.*, (1998) identified a kindred with three affected individuals, with early age of onset MDS, suggesting a possible inherited predisposition to this disease. Mandla *et al.*, examined whether band 5q31, a chromosomal region frequently associated with sporadic MDS, was involved in familial expression of MDS in this pedigree. Linkage analysis using polymorphic microsatellite DNA markers demonstrated that 5q31 did not cosegregate with MDS in this family. To date, only one family with familial MDS and a 5q- abnormality has been identified. Grimwade *et al.*, (1993) describe two sisters, both of whom had MDS and an interstitial deletion of 5q. The tracking of maternal and paternal polymorphisms in this study proved uninformative,

suggesting linkage analysis is not a viable strategy in the identification of the 5q- syndrome gene.

Novel genes identified from cloning chromosome translocations include the *BCL9* gene. Abnormalities of chromosome 1q21 are common in B-cell malignancies, and the nature of the involved gene(s) remains unknown (Willis *et al.*, 1998). A cell-line from a patient with B-cell ALL, which exhibited a t(1;14)(q21;q32), was used to identify the gene involved in the translocation. Novel gene *BCL9* was identified from sequencing full-length cDNA clones obtained from a normal human foetal brain cDNA library. Willis *et al.*, suggested that *BCL9* may be the target of translocation in some B-cell malignancies with abnormalities of 1q21, and that deregulated *BCL9* expression may be important in their pathogenesis. More recently, novel gene *TCL6* (T-cell leukaemia/lymphoma) was isolated from the breakpoint cluster region on chromosome 14q32.1 (a region often involved in chromosomal translocations and inversions in T-cell lymphoproliferative diseases) (Saitou *et al.*, 2000). The *TCL6* gene was found to be expressed in T-cell leukaemia carrying a t(14;14)q11;q32.1) chromosome translocation. However, like other T-cell genes, namely *TML1* and *TCL1*, *TCL6* was not expressed in normal T-cells, suggesting this novel candidate gene may be involved in leukaemogenesis (Saitou *et al.*, 2000).

There are several reports of chromosome translocations involving 5q (including 5q31) in MDS and myeloid leukaemia. For example, Borkhardt *et al.*, (2000) isolated the human *GRAF* gene (for GTPase regulator associated with the focal adhesion kinase pp125 (FAK)) from its fusion with the mixed-lineage leukaemia (*MLL*) gene in a unique t(5;11)(q31;q23) in an infant with juvenile myelomonocytic leukaemia, while Jaju *et al.*, (1999) describe a recurrent translocation at 5q35 , t(5:11)(q35;p15.5) in childhood AML. Moreover, the translocation at 5q33, t(5;12)(q33;p13) is a recurrent chromosomal abnormality in a subgroup of myeloid

malignancies including MDS. To date, there are no reported translocations breaking in the critical region of the 5q- syndrome at 5q31.3-q32.

Loss of heterozygosity at several chromosomal loci is a common feature of the malignant progression of human tumours. These regions are thought to harbour one or more putative tumour suppressor gene(s) playing a role in tumour development (Baffa *et al.*, 2000). Frequent LOH has been found at 17p in 3% to 4% of MDS and AML cases (Soenen *et al.*, 1998). Soenen *et al.*, found a strong correlation in AML and MDS between a 17p deletion and a typical form of dysgranulopoiesis containing pseudo-Pelger-Huët hypolobulation and the presence of small vacuoles in granulocytes. The authors also found a strong correlation between the 17p deletion and a *p53* mutation, suggesting that MDS and AML with a 17p deletion constitute a new morphologic-cytogenetic-molecular "entity" in those disorders ("17p- syndrome"). Soenen *et al.*, studied seventeen cases of AML and MDS with a 17p deletion. In 14/17 cases, FISH showed a 17p deletion of variable extent but that always included deletion of the *p53* gene. All fourteen patients had typical dysgranulopoiesis, and all but one had *p53* mutation and/or overexpression. These findings reinforce the morphologic, cytogenetic, and molecular correlation found in the 17p- syndrome and suggest a pathogenetic role for inactivation of tumour suppressor gene(s) located in 17p, especially the *p53* gene.

The identification of other novel coding sequences that may be implicated in cancer and disease has relied on a number of different molecular techniques over the last decade. These include screening zoo blots that were used to discover four novel genes in the class II region of the human major histocompatability complex (Hanson *et al.*, 1991); and using YACs for hybridisation against cDNA libraries. More recently, the strategies of exon trapping and direct selection have been used. However, more rapid process has been made by the screening of EST databases.

4.1.2 Techniques for isolating novel coding sequences

4.1.2.1 Exon trapping

Exon trapping is a technique that exploits mRNA splicing to discover genes directly from genomic DNA capturing complete internal exons (Buckler *et al.*, 1991), or 3'-terminal exons (Krizman *et al.*, 1993). It has been successfully used to isolate the novel Huntington's disease (*HD*) gene (The Huntington's Disease Collaborative Research Group, 1993). Exon trapping was also the primary technique used to search for the tumour suppressor gene(s) on chromosome 5q that may play an important role in the progression of lung cancer (Hosoe, 1996).

4.1.2.2 Direct selection

Direct selection is an expression-based gene identification technique that can rapidly identify cDNAs within large genomic regions (Del Mastro and Lovett, 1997). The technique involves the hybridisation of a cDNA library to a genomic clone. cDNAs homologous to the target are selected and subsequently enriched by PCR amplification. As with exon trapping, direct selection has been used successfully to isolate novel coding sequences associated with disease. One study isolated five novel genes from the cri-du-chat critical region at 5p15.2 using cosmids from the LANL chromosome 5 specific cosmid library (Simmons *et al.*, 1995).

4.1.2.3 YAC hybridisation

The YAC hybridisation method involves the hybridisation of a radiolabeled YAC, containing the genomic DNA of interest, to a cDNA library that has been transformed into bacteria and replicated onto nitrocellulose filters (Fidler and Boultwood, 1997). A number of groups have successfully used this technique to

isolate novel coding sequences. Wallace *et al.*, (1990) used it to identify part of the *NF1* gene, and Snell *et al.*, (1993) isolated seven novel cDNAs mapping to the candidate region of the Huntington's gene. The YAC hybridisation method was also the primary technique used to isolate novel coding sequences from the critical region of the 5q- syndrome (Boultwood *et al.*, 1997).

All the above methods, although successful in identifying novel coding sequences, have not made a major impact on genome research. The turning point came recently with the arrival of ESTs that were conceived as a shortcut to the finish line (Adams *et al.*, 1991; Boguski, 1995). ESTs are 5' and 3' terminal sequence reads from cDNA clones that are thought to represent nearly all genes in the human genome as well as other species, e.g. mouse, *Drosophila*, and *Caenorhabditis elegans*.

4.1.3 Database searching and ESTs

Database searching is now the primary technique for the isolation of novel coding sequences. The field of EST research began tentatively in 1991 with the large-scale sequencing project being undertaken predominantly by the private sector. The public data collection was boosted with the launch of the Washington University Human EST project in October, 1994. This had a goal to provide up to 400,000 ESTs for the public domain by March 31, 1996 (Boguski and Schuer, 1995). This number has increased dramatically since then. The number of public entries on January 12, 2001 was 6,994,862, of which 2,953,517 were human. These ESTs are available to the public in the form of their I.M.A.G.E. cDNA clones. The clones are from three hundred and sixty different human cDNA libraries and available free of any royalties. Over 3.8 million distinct cDNA clones are now arrayed; from which over 2.3 million 5' and/or 3' sequences have been deposited into db(EST). A variety of methods all indicate that the I.M.A.G.E. collection is likely to represent over 60,000 distinct human genes at this time (I.M.A.G.E. Consortium home page, January, 2001).

Due to this rapid progress, we decided to use I.M.A.G.E. cDNA clones derived from ESTs as our primary resource to identify novel coding sequences mapping to the critical region of the 5q- syndrome.

4.1.4 Isolating novel genes mapping to the critical region of the 5q- syndrome

The Chromosome 5 Human Gene Maps at NCBI (GeneMap'98 and '99) have identified approximately one hundred and fifty cDNA sequences, of which forty-six are represented by known genes, and one hundred and four are represented as ESTs, mapping to the interval between DNA markers D5S410 and D5S487. Each interval on the GeneMap illustrates the number of known genes and unidentified transcripts (ESTs), and represents their position relative to each other on the transcript map. The GeneMaps therefore served as the primary source for selecting ESTs within the critical region of the 5q- syndrome. Despite the use of normalised libraries to produce these ESTs, there is a considerable degree of redundancy in the data. To overcome this, the UniGene set was created as another source for selecting ESTs. The UniGene set comprises a non-redundant set of unique human 3' UTRs that are divided into clusters of sequences that are most likely to be derived from the same gene. In January, 2001, there were over one thousand separate entries of ESTs for human chromosome 5 at UniGene, each represented by at least one sequence, with some transcripts represented by ninety-nine sequences.

We selected ESTs from both the Human chromosome 5 GeneMaps and from UniGene (as unidentified transcripts), with the aim of identifying novel coding sequences mapping to the critical region of the 5q- syndrome. I.M.A.G.E. cDNA clones from which these ESTs were derived were used for analysis. A third source of ESTs came from the collaboration with Professor Charles Auffray at Genethon. Auffray *et al.*, (1995) derived 26,938 ESTs from skeletal muscle and infant brain cDNA clones. Of these ESTs, two thousand five hundred were assigned to human

chromosomes, one hundred and thirty of these binned to chromosome 5, and six ESTs localised to the critical region of the 5q- syndrome at 5q31-q33. In total, twenty-three cDNAs were included in this study to identify the putative tumour suppressor gene associated with the 5q- syndrome.

4.2 Materials and Methods

4.2.1 EST identification

A collaboration with Professor Charles Auffray at Genethon was established to identify ESTs mapping to the YAC contig spanning the approximate 5Mb critical region of the 5q- syndrome at 5q31-q33, flanked by the genes *FGF1* and *IL12ß*. The ESTs were localised by PCR amplification. Two ESTs were isolated from the Stratagene skeletal muscle cDNA library, and 4 ESTs were isolated from the normalised infant brain cDNA library.

Following the reduction of the 5q- syndrome critical region to approximately 3Mb, the Human Chromosome 5 GeneMap'98 was accessed for ESTs between the DNA markers D5S410 and D5S487 at 5q31.3-q33, flanked by the genes *ADRß2* and *IL12ß*. ESTs were selected following certain criteria; they were mapped independently by more than one group, and were expressed in haematological tissues. Five transcripts were identified matching the aforementioned criteria. Two ESTs representing each transcript were selected for further analysis.

The updated Human Chromosome 5 GeneMap'99 was also accessed for ESTs between the DNA markers D5S410 and D5S487 which span the critical region of the 5q- syndrome at 5q31.3-q33. Twelve transcripts were identified from the GeneMap'99. Two ESTs representing each transcript were selected for further analysis.

In total, twenty-three transcripts were identified from Genethon, the GeneMap'98, and GeneMap'99, and selected for further analysis.

4.2.2 I.M.A.G.E. cDNA clones

I.M.A.G.E. cDNA clones from the ESTs of which they were derived, were obtained from Professor Charles Auffray at Genethon or the UK HGMP-RC as stabs in agar. Single colonies were obtained by plating onto LB ampicillin (50mg/ml) plates. A single colony was then inoculated into a 10ml LB culture containing ampicillin. Plasmid DNA was obtained using the QIAprep® Spin Miniprep Kit. The insert was excised using the appropriate restriction enzymes and purified for use as a probe with the Wizard® PCR Preps DNA Purification System.

4.2.3 Samples

Ten patients with the classical features of the 5q- syndrome, including a 5q deletion as the sole karyotypic abnormality were included in the study. Granulocyte and mononuclear cells were separated from 40mls of peripheral blood by ficoll gradient centrifugation (Boyum, 1984). The granulocytes showed a high level of purity (≥95%). Mononuclear cells (specifically T-lymphocytes) were isolated by erythrocyte rosetting and showed a purity of ≥90%. High molecular weight DNA was obtained from the fractionated blood leukocytes by Nucleon® extraction. Granulocyte DNA fractions from the peripheral blood of healthy individuals were used as controls. High molecular weight DNA was obtained from a human/mouse hybrid cell line with human chromosome 5 as its only human complement.

4.2.4 Gene dosage analysis

Gene dosage analysis was used to confirm the localisation of the EST (cDNA clone) to chromosome 5; and to determine the loss or retention of the clone in the patient granulocyte DNA (**Chapter 3 section 3.2.4**). Gene dosage experiments were carried out on at least two separate occasions.

4.2.5 Northern analysis

cDNA clones which hybridised to a single fragment in hybrid 5 DNA and showed a 50% dosage reduction were hybridised to Multiple Tissue Northern (MTN) blots (Clontech) (**Chapter 3 section 3.2.5 and Table 3.1**).

4.2.6 Southern analysis

cDNA clones which hybridised to a single fragment in hybrid 5 DNA and showed a 50% dosage reduction were hybridised to Southern blots to screen for rearrangements (**Chapter 3 section 3.2.6**). Granulocyte DNA fractions from eight 5q- syndrome patients and DNA from control samples were digested with restriction enzymes *Eco*RI, *Pst*I, *Hind*III, *Bgl*II, *Pvu*II, and *Eco*RV. Southern blot filters were prepared and hybridised separately with the three novel cDNAs.

4.2.7 Direct sequencing

I.M.A.G.E. cDNA clones representing each novel cDNA were sequenced as either single-stranded or double-stranded templates, as previously described, by the dideoxy chain termination method (Sanger *et al.*, 1977) (**Chapter 2 section 2.12.1**). Clones were sequenced using the Cy5 Autoread sequencing kit (Amersham Pharmacia Biotech) (**Chapter 3 section 3.2.7**). Each I.M.A.G.E. cDNA clone was sequenced in full and then subjected to a GenBank search for homology with known genes and overlapping clones to generate the full-length cDNA.

4.2.8 Overlapping cDNA clones

Sequence data from each I.M.A.G.E. cDNA clone was subjected to a homology search against the EST database db(EST) at NCBI for overlapping cDNA clones to generate the full-length cDNA (**Chapter 3 section 3.2.8**). Sequence data from the

overlapping clone was added to the sequence from the I.M.A.G.E. cDNA clone, and the 'new' sequence submitted to db(EST).

4.2.9 cDNA library screening

If no overlapping clones were identified from db(EST) or UniGene, the cDNA clone insert was screened against cDNA libraries. In the first instance, a foetal brain cDNA library was selected as this tissue expresses a wide variety of genes. Seven high-density gridded cDNA filters were used in the study. In addition, a collaboration with RZPD was established (**Chapter 3 section 3.2.9**). Positive clones were sequenced and the 'new' sequence submitted to db(EST).

4.2.10 Dot blot analysis

If more than 5 'positive' clones were identified from screening the foetal brain cDNA library filters, the clones were first analysed by Dot blot analysis. Dot blot analysis was used to select strongly positive clones from those identified from screening the foetal brain cDNA library filters.

2cm x 2cm squares were drawn on Hybond N^+ membrane to form a grid. Each square represents each cDNA clone to be tested. Two squares are added on to the bottom of the grid for the positive and negative controls. Each cDNA clone to be tested is spotted onto the membrane to form a circle approximately 0.7cm in diameter. The original cDNA clone acts as the positive control. Genomic DNA from another region on chromosome 5, e.g. from a YAC which does not contain the cDNA sequence was used as the negative control.

Following autoradiography, the signal intensity of each cDNA clone was compared with the signal intensity of the positive control.

4.2.11 RACE PCR

The technology of RACE PCR was used to generate 'new' sequence when no overlapping clones were identified from screening db(EST) and cDNA libraries (**Chapter 2 section 2.13**). The libraries chosen were tissue-specific to the gene of interest.

1. Gene-Specific Primers were designed from the 5' and/or 3' end of the cDNA of interest dependent on whether 5' or 3' RACE was to be performed. Details of the primers, including their melting temperature (Tm), and choice of thermal profile are shown in Table 4.1.

2. The Marathon-Ready™ cDNA templates used in the 25µl RACE PCR reaction included; human foetal liver, lung, bone marrow, pituitary gland, small intestine, foetal skeletal muscle, testis, foetal brain, foetal spleen, hypothalamus, and foetal thymus.

3. The RACE PCR products were subcloned and prepared for sequencing as previously described (**Chapter 2 section 2.14**).

Table 4.1 **RACE PCR primer conditions for novel cDNAs A3B02, 43911, and 199067**

I.M.A.G.E. cDNA	5' or 3' RACE	Primer name	GSP primer sequence 5'-3'	Tm of primer	Touchdown/ three-step PCR
A3B02	3'	A3B02R3 (GSP1)	AATCTGGGACTTGAGACCTG	61°C	Three-step
		A3B02R1 (GSP2)	GGTAGCTGGAGACTTCCCAT	62°C	
	5'	A3B02F10 (GSP1)	TTTCCTGCCACATCTGCTCTCCAT	72°C	Touchdown
		A3B02F16 (GSP2)	CATTTCCTGCCACATCTGCTCTCC	72°C	
43911	3'	43911F1 (GSP1)	CCAATCATCATGATACATAAGATAAGT	60°C	Three-step
		43911F3 (GSP2)	GGCAATGAAGGGATATGTTTTTAGACT	66°C	
	5'	43911R6 (GSP1)	CATGCATATTTGTGAAGAAACACCCTT	68°C	Three-step
		43911R4 (GSP2)	CAGCTTTGGAAACTTAGGCTAAGTTA	64°C	
	5'	43911R12 (GSP1)	GGCAGCCATTATGGCAATGAAGGG	74°C	Touchdown
		43911R10 (GSP2)	GCCATTCCAAGGAACACCCATCC	73°C	
199067	5'	199067R3 (GSP1)	GACTTGAGAGCAAGAGTGGCCTG	64°C	Three-step
		199067R1 (GSP1)	CCCTGGGCCCCTGCTAAGAAATC	66°C	
	5'	199067R4 (GSP1)	GTTGGTTAGTGAATGACTCCTGCTC	63°C	Three-step
		199067R6 (GSP1)	CTGATTGCTACTGACCCAACCAAC	63°C	
	5'	199067R7 (GSP1)	CACAGTAAGTCCTCCCTTGTCGTC	64°C	Three-step
		199067R9 (GSP2)	CGCTATACATACATTTAATGTATTGCAG	59°C	
	5'	199067R9 (GSP2)	CGCTATACATACATTTAATGTATTGCAG	59°C	Three-step
		199067R11 (GSP2)	AATAACAGTATTTTGAGAAAATGCTG	55°C	

4.2.12 Localisation to the YAC contig

Each cDNA was sublocalised by PCR screening to the YAC contig encompassing the critical region of the 5q- syndrome (Kostrzewa *et al.*, 1998) as previously described (**Chapter 2 section 2.15**). PCR primer pairs were designed from the cDNA sequence. Details of the primer conditions are shown in Table 4.2.

4.2.13 Expression analysis

cDNA clones were analysed by RT-PCR analysis for expression in RNA extracted from $CD34^+$ cells, as previously described (**Chapter 2 section 2.17**)

4.2.14 Database analysis using the Genetics Computer Group (GCG) software package

The sequence generated from each novel cDNA clone was subjected to a FastA nucleotide, and BLAST protein search as previously described (**Chapter 3 sections 3.2.12.1, 3.2.12.2**). The sequence was then submitted for submission at GenBank as previously described (**Chapter 3 section 3.2.12.6**).

Table 4.2 YAC localisation PCR conditions for novel cDNAs 43911, 195312, 195971, 120101, and 199067

I.M.A.G.E. cDNA	Primer name	Primer sequence 5'-3'	Tm of primers	PCR product size	Positive YAC(s)
43911	43911F2 43911R2	AGCTGGTGGTCTACTTTATC TTCTGGCCTAACAGCAGAC	60°C	103bp	816D6 (C)*
195312	195312F1 195312R1	CCCTCAACATTCAATCCC TCTAAAGCATTGTTTCTGCC	60°C	123bp	816D6 (C)*
195971	195971F1 195971R2	TGCATCTCTCTTCAAAGATCAC CAATAGAAAAACTCCAGGTGAC	60°C	106bp	757H2 + 914A10 (overlapping)
120101	120101F1 120101F1	GACCTCACAGAAGTAAACCC ATCATAAGGCAAAGGCAG	60°C	134bp	757H2 + 914A10 (overlapping)
199067	199067F1 199067R1	GCCACTCTTGCTCTCAAGTC CGGGGCATGCTCTTAACAG	60°C	143bp	816D6 (C)*

* (C) denotes chimaeric YAC

NB. A3B02 had previously been localised to YAC 816D6 by Professor Charles Auffray at Genethon.

4.3 Results

4.3.1 EST identification

Six ESTs from the collaboration with Professor Charles Auffray at Genethon, were localised to the YAC contig mapping to the approximate 5Mb critical region at 5q31-q33, by PCR amplification, see Table 4.3. A further 17 novel cDNAs were identified from the Human Chromosome 5 GeneMaps'98 and '99, and the UniGene set, see Tables 4.3 and 4.4. I.M.A.G.E. cDNA clones from which each EST was originally derived were obtained.

4.3.2 Gene dosage analysis

Gene dosage analysis was carried out on 12/23 novel cDNAs, see Figure 4.1. Ten out of twelve (83%) cDNA clones were shown to map to the critical region of the 5q- syndrome at 5q31.3-q33. Probe 1ja10 was shown to hybridise to 4 fragments in the granulocyte DNA from the patients and controls, but 1 fragment in Hybrid 5. This suggested cDNA 1ja10 to be a recombinant gene or a member of a gene family. Direct sequencing of cDNA 1ja10 as a single-stranded template generated 1344bp of sequence. The 365bp EST sequence was not found within the cDNA as expected, suggesting the clone was incorrect. cDNA 1ja10 was discarded from this point.

The remaining 9 novel cDNAs were selected for further analysis, see Figure 4.2.

Table 4.3 ESTs identified from the collaboration with Genethon, and the Human Chromosome 5 GeneMap'98

I.M.A.G.E. clone name	GenBank Accession No.	Source	D marker interval	Tissue source (cDNA library)	I.M.A.G.E. clone insert size
A3B02	AF010242	Genethon	D5S1580	Stratagene skeletal muscle	1216bp
Cdy-17a06	AF010244	Genethon	D5S1688	Normalised infant brain	783bp
Cda-19c10	AF010245	Genethon	Genethon 5q31-q33	Normalised infant brain	347bp
Cda-1ja10	AF010243	Genethon	D5S1652	Normalised infant brain	1344bp
Cda-1jh07	Z43753	Genethon	Genethon 5q31-q33	Normalised infant brain	1500bp
Bda-87b11	Z28741	Genethon	Genethon 5q31-q33	Stratagene skeletal muscle	1232bp
43911	AF156165	GeneMap'98	D5S470	Soares infant brain 1NIB	1497bp
195312	R92031	GeneMap'98	D5S470-D5S410	Soares foetal liver spleen 1NFLS	1000bp
195971	R91397	GeneMap'98	D5S410-D5S487	Soares foetal liver spleen 1NFLS	719bp
120101	AF156166	GeneMap'98	D5S410-D5S487	Soares foetal liver spleen 1NFLS	1269bp
199067	H82831	GeneMap'98	D5S410-D5S487	Soares foetal liver spleen 1NFLS	1400bp

Table 4.4 ESTs identified from the Human Chromosome 5 GeneMap '99

I.M.A.G.E. clone name	GenBank Accession No.	D marker interval	Tissue source (cDNA library)	I.M.A.G.E. clone insert size
197258	R86964	D5S402-D5S2090	Soares foetal liver spleen 1NFLS	1067bp
110211	T71275	D5S436-D5S413	Soares foetal liver spleen 1NFLS	1105bp
485953	AA040631	D5S410-D5S487	Soares pregnant uterus NbHPU	787bp
192250	H41167	D5S470-D5S410	Soares foetal liver spleen 1NFLS	1627bp
30879	R41775	D5S410-D5S487	Soares infant brain 1NIB	2007bp
277516	N56962	D5S434-D5S2013	Soares multiple sclerosis 2NbHMSP	550bp
327361	W02135	D5S410-D5S487	Soares foetal heart NbHH19W	623bp
141271	R67401	D5S470-D5S410	Soares placenta Nb2HP	1023bp
265726	N22851	D5S470-D5S410	Soares melanocyte 2NbHM	1743bp
341099	W58211	D5S410-D5S487	Soares foetal heart NbHH19W	490bp
143772	R76720	D5S410	Soares pregnant uterus NbHPU	1187bp
240080	H82404	D5S410-D5S487	Soares foetal liver spleen 1NFLS	2344bp

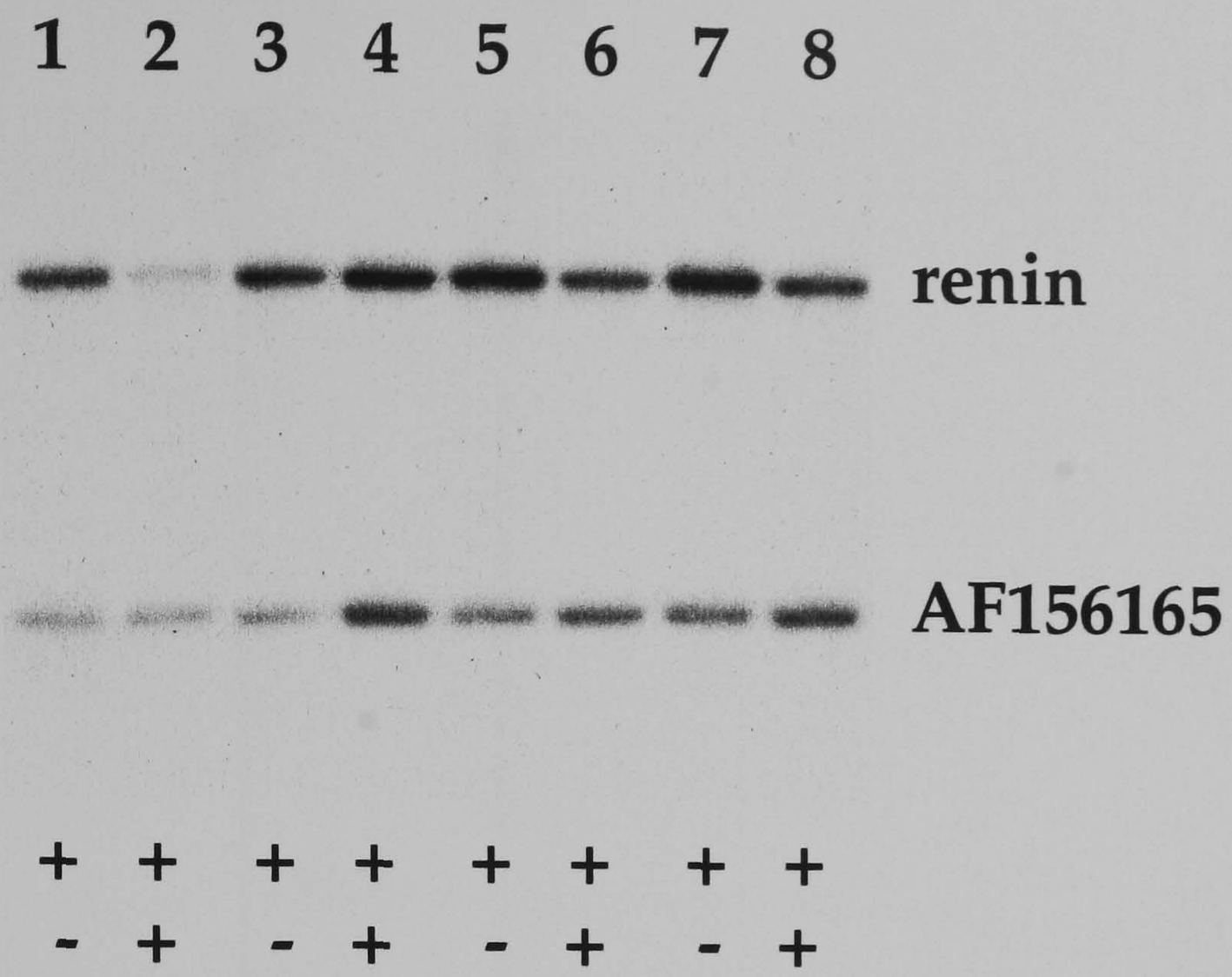


Figure 4.1

Representative gene dosage analysis of novel cDNA AF156165. DNA obtained from the granulocyte fractions of 4 patients (lanes **1**, **3**, **5** and **7**) and healthy controls (lanes **2**, **4**, **6** and **8**) was digested with *Eco*RI and simultaneously hybridised to a probe for AF156165 and a probe for the *renin* gene. ++ indicates the presence of two copies of the AF156165 gene and + - indicates the deletion of one copy of the gene.

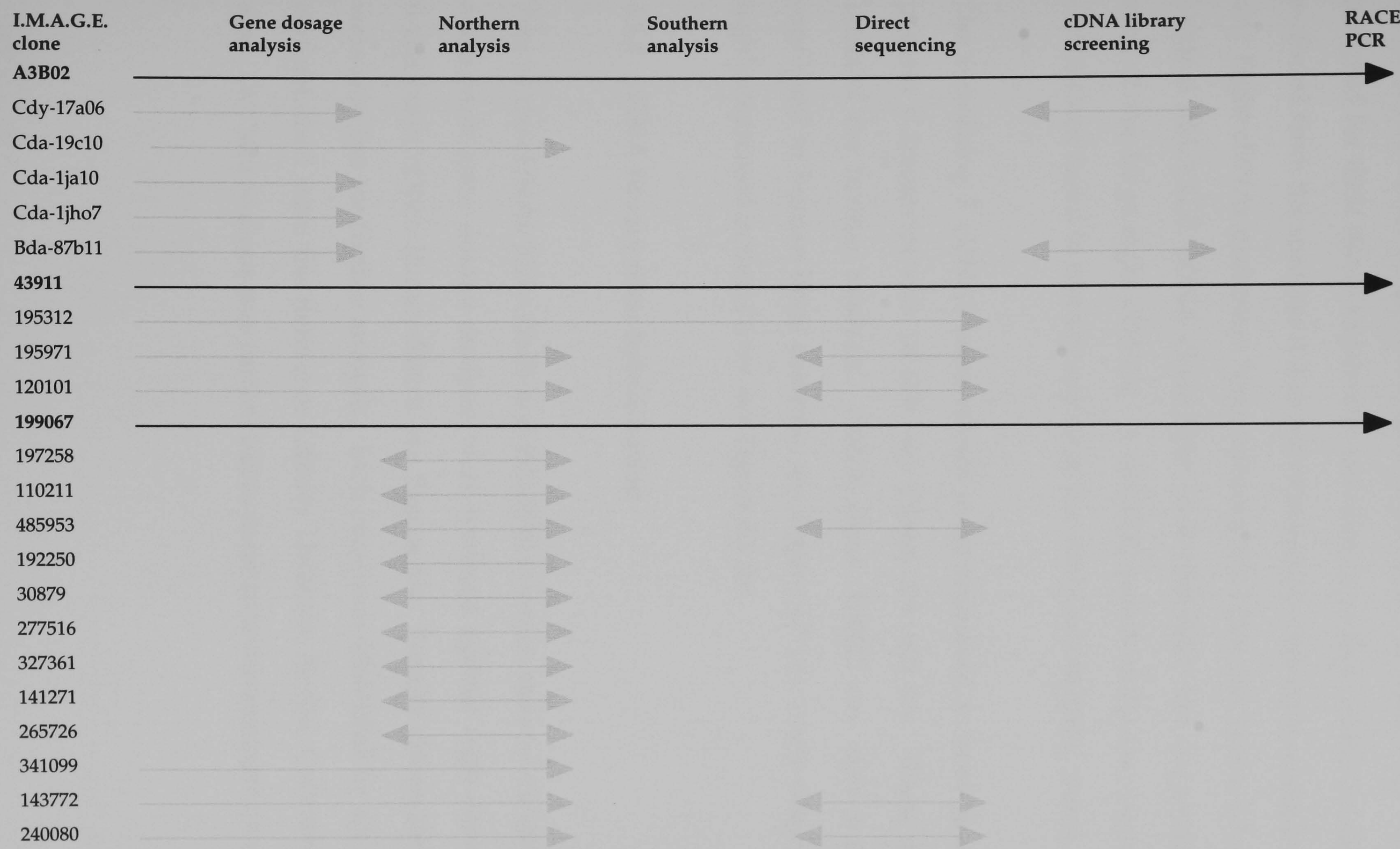


Figure 4.2 Data map of novel ESTs identified from Genethon, the Human Chromosome 5 GeneMaps '98 and '99, and the UniGene set. The horizontal lines show the analysis covered for each novel EST. The arrows show where the analysis stopped for each novel EST. The map shows complete molecular analysis for three novel ESTs (boldface).

4.3.3 Northern analysis

The 9 cDNAs from the gene dosage analysis plus the remaining 11 cDNAs were analysed for their tissue expression and transcript size. cDNA clone 197258 was excluded from the study as it was not expressed in human bone marrow, see Table 4.6. Eight cDNAs possessed large transcripts ≥7.5kb, see Tables 4.5 and 4.6. The original I.M.A.G.E. cDNA clone from each transcript was sequenced in full for each of the large eight cDNAs. A db(EST) search using the sequence of each cDNA clone failed to identify any large (i.e. ≥1kb) overlapping clones.

The remaining 7 cDNAs were shown to be expressed in bone marrow and to possess a transcript size ≤4.4kb, see Tables 4.5 and 4.6. These cDNAs were selected for further analysis. cDNA clone A3B02 was shown to be highly expressed in human bone marrow, see Figure 4.3 (a), while clone 248808 was highly expressed in foetal liver, see Figure 4.3 (b).

4.3.4 cDNA library filter hybridisation

Two of the cDNAs (Cdy-17a06 and Bda-87b11) which showed the deletion of one allele from gene dosage analysis were screened against high-density gridded, foetal brain cDNA library filters, see Figure 4.4. Twenty clones for each cDNA were selected for further analysis. Each clone was identified by x,y co-ordinates and obtained from the Reference Library Database, Berlin, Germany as stabs in agar. Dot blot analysis was carried out to determine the true positive clones.

Table 4.5 **Expression patterns and transcript sizes (kb) of novel cDNAs mapping to the critical region of the 5q- syndrome, identified from Genethon and the Human Chromosome 5 GeneMap'98**

	MTN 1 (size in kb)								MTN 2 (size in kb)					
Novel cDNA	ht	br	pl	lu	li	sm	ki	pa	s	ln	t	pbl	bm	fl
A3B02	4.4					4.4							4.4	
43911	4.4	4.4	4.4	4.4	4.4	4.4	4.4	4.4	4.4	4.4	4.4	4.4	4.4	4.4
195312	10	10	10	10	10	10	10	10	10	10	10	10	10	10
120101	-	-	-	-	-	-	-	-	PCR+	PCR+	PCR+	PCR+	PCR+	PCR+
195971	-	-	-	-	-	-	-	-	PCR+	PCR+	PCR+	PCR+	PCR+	PCR+
199067	-	-	-	-	-	-	-	-	1.9	1.9	1.9		1.9	1.9

ht - heart, br - brain, pl - placenta, lu - lung, li - liver, sm - skeletal muscle, ki - kidney, pa - pancreas, s - spleen, ln - lymph node, t - thymus, pbl - peripheral blood leukocytes, bm - bone marrow, fl - foetal liver

PCR+ indicates that the cDNA clone was expressed according to RT-PCR analysis

Table 4.6 **Expression patterns and transcript sizes (kb) of novel cDNAs mapping to the critical region of the 5q- syndrome, identified from the Human Chromosome 5 GeneMap'99**

	MTN 2 (size in kb)					
Novel cDNA	**s**	**ln**	**t**	**pbl**	**bm**	**fl**
265726	9.5	9.5	9.5	9.5	9.5	9.5
141271	7.5	7.5	7.5	7.5#	7.5	7.5#
327361	>10.0	>10.0	>10.0	>10.0	>10.0	>10.0
277516	7.5	7.5	7.5	7.5	7.5	7.5
30879	7.5	7.5	7.5	7.5	7.5	7.5
110211	4.4, 8.5	4.4, 8.5	4.4, 8.5	4.4, 8.5	4.4, 8.5	4.4, 8.5
192250	4.4, 7.5	4.4, 7.5	4.4, 7.5	4.4, 7.5	4.4, 7.5	4.4, 7.5
197258	smear	smear	smear	smear	smear	smear
485953	2.5	2.5	2.5	2.5	2.5	2.5
240080	2.4	2.4	2.4		2.4	2.4*
341099	3.5, 4.5	3.5, 4.5	3.5, 4.5	3.5, 4.5	3.5, 4.5	3.5, 4.5
143772	3.0, 4.4	3.0, 4.4	3.0, 4.4	3.0, 4.4	3.0, 4.4	3.0, 4.4

s - spleen, ln - lymph node, t - thymus, pbl - peripheral blood leukocytes, bm - bone marrow, fl - foetal liver

very weak signal

* very high level of expression

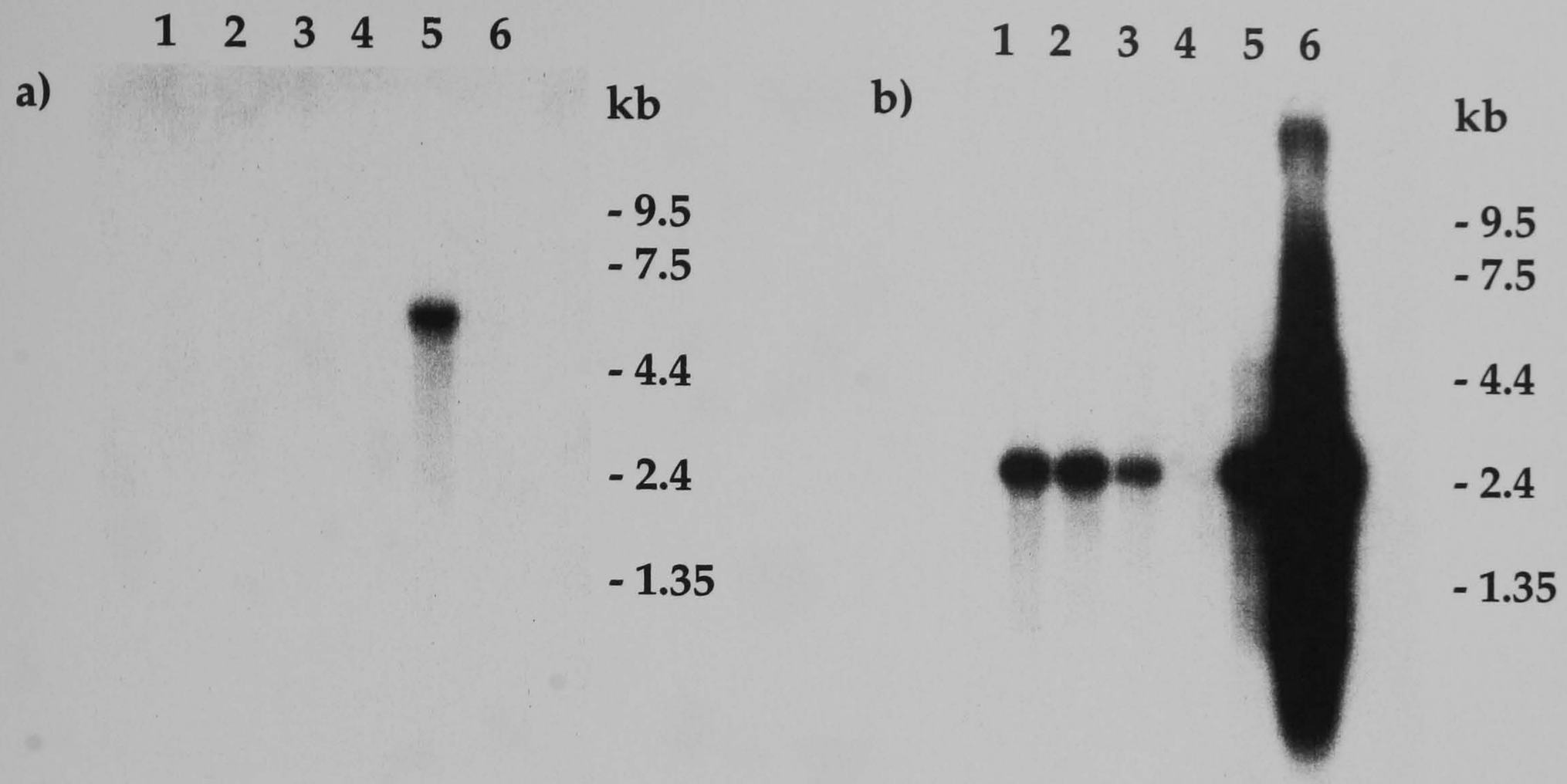


Figure 4.3

Representative Northern blot analysis of (a) I.M.A.G.E. cDNA A3B02, and (b) I.M.A.G.E. cDNA 240080. The MTN blots included 2µg of poly (A^+) RNA from; spleen (1), lymph node (2), thymus (3), peripheral blood leukocytes (4), bone marrow (5), and foetal liver (6). Sizes of RNA marker bands (kb) are indicated approximately.

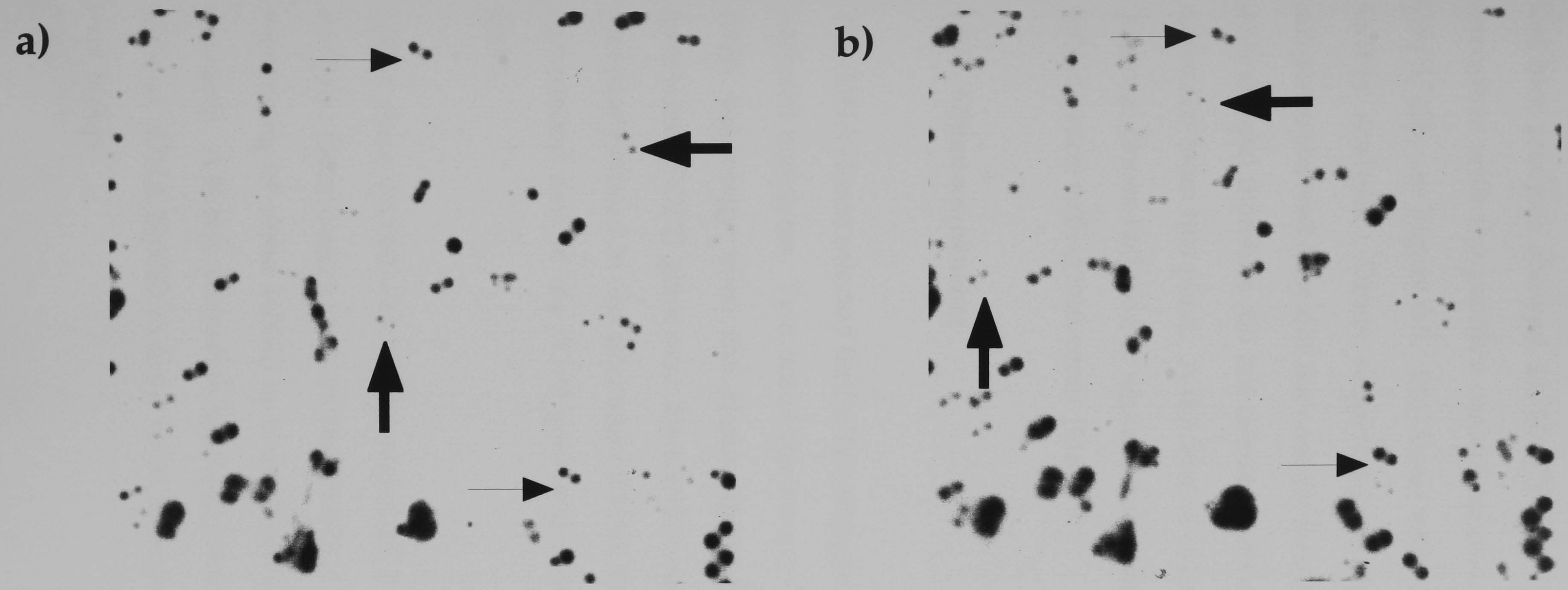


Figure 4.4

Representative results obtained from hybridisation of high-density gridded, foetal brain cDNA library filters, with cDNA probes illustrating the similar hybridisation pattern obtained from different cDNA probes. (a) Filter hybridised with cDNA probe Cdy-17a06, (b) same filter after stripping hybridised with cDNA probe Bda-87b11. Small arrows indicate positive clones identified with both cDNA probes. Large arrows indicate positive clones unique to each cDNA.

4.3.5 Dot blot analysis

Dot blot analysis showed 14/20 (70%) of 'positive' clones had similar signal intensities with the positive control following hybridisation with cDNA probe Cdy-17a06, see Figure 4.5. Six clones with the largest inserts were selected for further analysis. Direct sequencing of these 'positives' as single-stranded templates showed they did not contain their EST sequence from which they were derived, and therefore did not overlap with clone Cdy-17a06. These clones were discarded from this point. A dot blot of clone Bda-87b11 showed 14/20 (70%) of 'positive' clones had similar signal intensities with the positive control following hybridisation with cDNA probe Bda-87b11.

4.3.6 Direct sequencing

The cDNA clones selected for further analysis were sequenced as single or double-stranded templates. Two out of the seven cDNAs (4885953 and 143772) showed 100% homology with the *Homo sapiens* CCR4-associated factor 1 *CNOT8* (previously *POP2*) gene over their entire sequence. The *POP2* gene had been previously cloned in our laboratory (Fidler *et al.*, 1999). These two ESTs therefore represented part of the *POP2* gene and were discarded from the study at this point.

cDNA clone 240080 was highly expressed in foetal liver according to Northern analysis. Gene dosage analysis had shown it not to map to chromosome 5. Direct sequencing of clone 240080 as a double-stranded template generated 478bp of sequence. A FastA nucleotide homology search utilising the GenEMBL databases showed cDNA 240080 to have 100% homology with the *Homo sapiens H19* gene over 416bp.

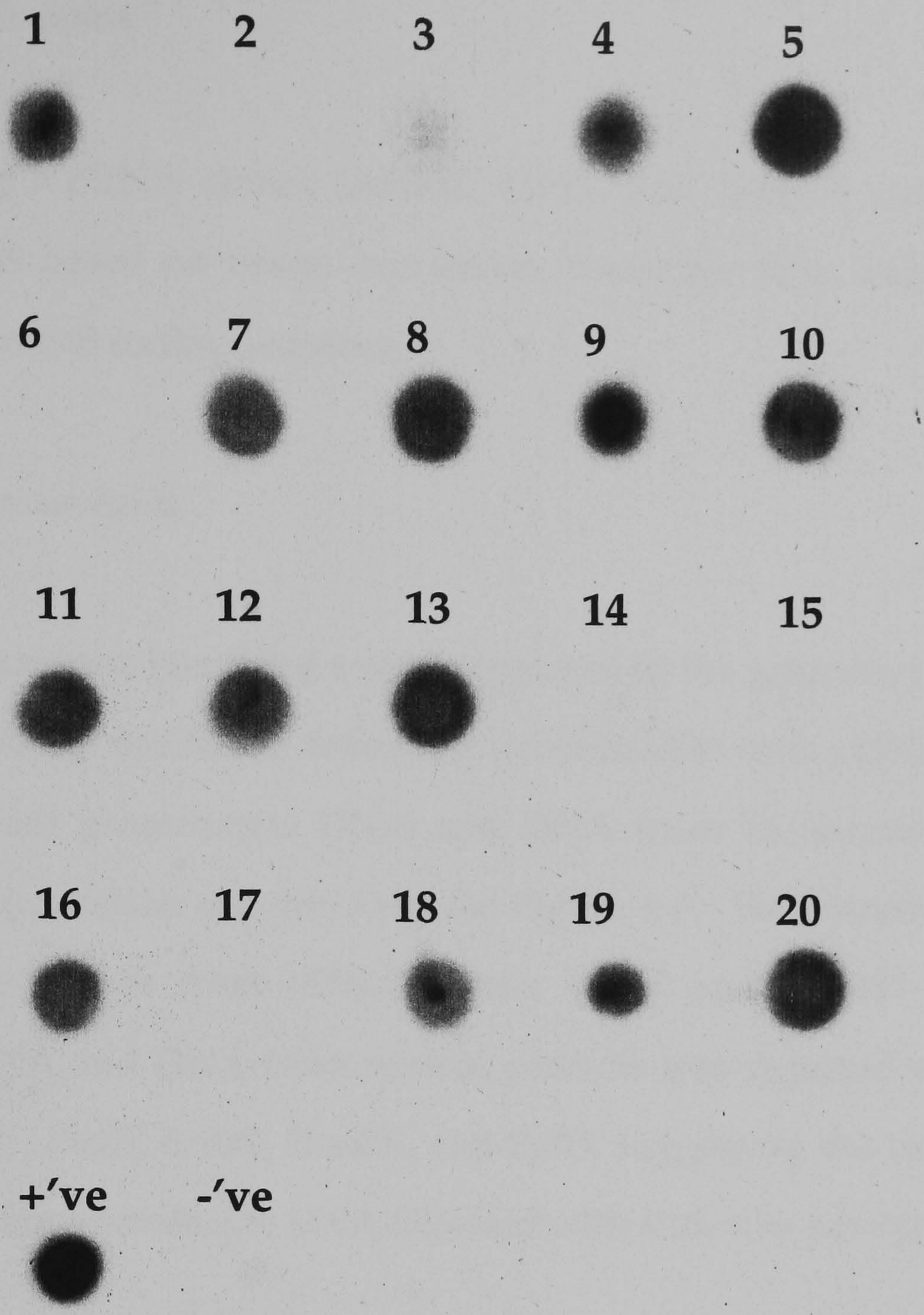


Figure 4.5

Representative results from dot blot analysis following hybridisation with cDNA probe Cdy-17a06. The true positives are shown as strong black signals. The false positives are the faint or non-existent signals. YAC 176D01 (positive for clone Cdy-17a06) was used as the positive control. YAC 15DB10 (from the opposite end of the contig) was used as the negative control.

This result confirmed both the Southern and Northern data, and the EST was discarded at this point.

The remaining 3 cDNA clones (A3B02, 43911, and 199067) were selected for further analysis based on tissue expression, transcript size, and probability of completing their full coding sequence.

4.3.7 Southern analysis

Southern blot analysis identified a rearrangement in the granulocyte DNA of one patient with the 5q- syndrome following hybridisation with cDNA clone A3B02 when the patient granulocyte DNA and DNA from 18 normal controls was digested with restriction enzyme *Pst*I, see Figure 4.6. No rearrangements were seen with the probes from cDNA clones 43911 and 199067 when patient granulocyte DNA and DNA from normal controls was digested with restriction enzymes *Eco*RV, *Pvu*II, *Eco*RI, *Hind*III, and *Bgl*II, suggesting the rearranged band seen in the patient granulocyte DNA digested with *Pst*I, was a polymorphism.

4.3.8 Overlapping cDNA clones

A GenBank homology search on the sequence of cDNA clones A3B02, 43911 and 199067 showed the cDNAs to be completely novel. A db(EST) search using the sequence of each cDNA clone identified several overlapping clones. Direct sequencing of approximately 60% of these overlapping clones generated sequence that overlapped with the cDNA clone sequence from which it was obtained with 100% homology over part of the cDNA. The remaining 40% of clones either did not contain their EST sequence and/or did not overlap with their respective cDNA.

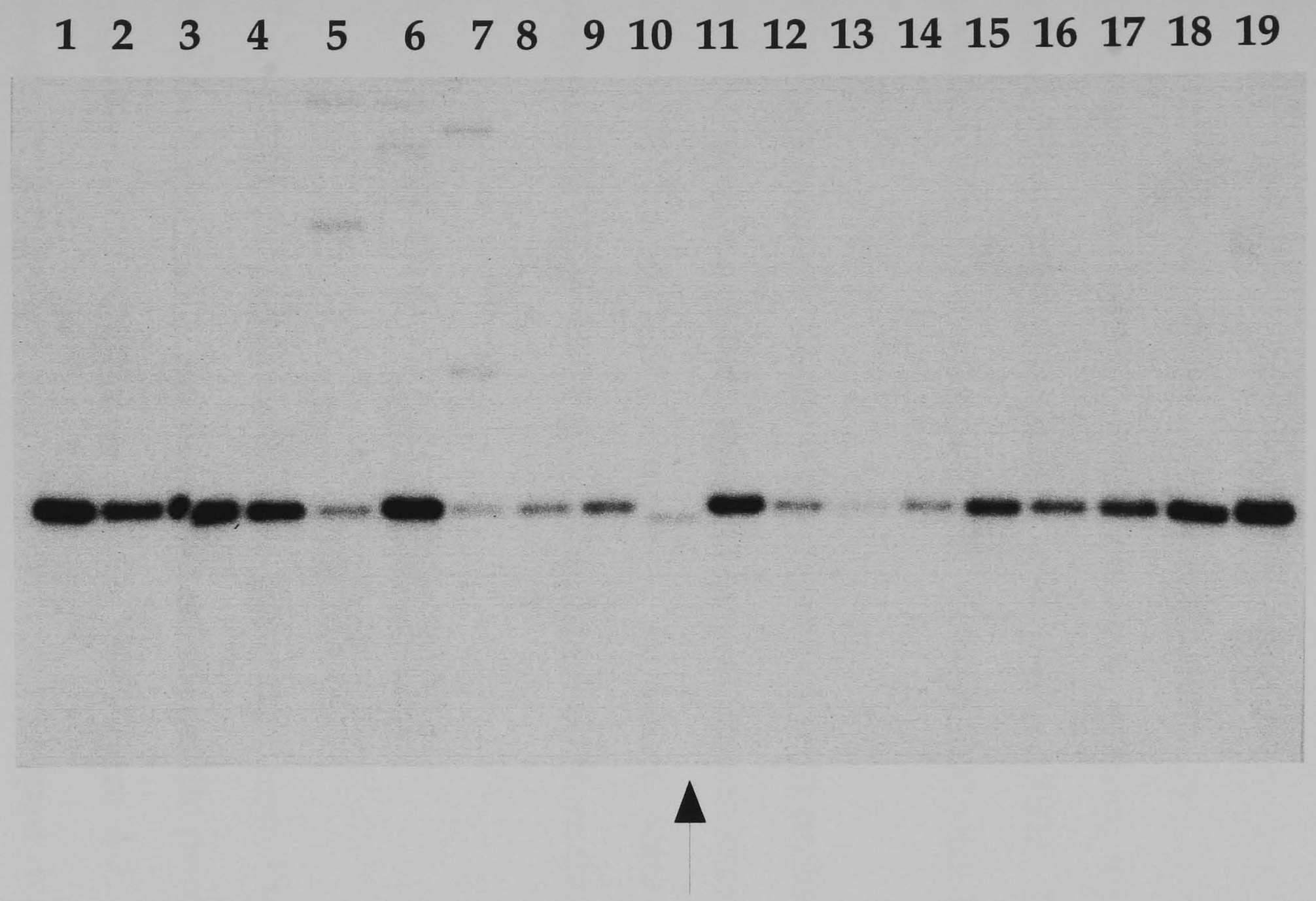


Figure 4.6

Representative Southern blot analysis of I.M.A.G.E. cDNA clone A3B02. DNA obtained from the granulocyte fraction of one patient (10) and the peripheral blood of eighteen healthy controls (C) was digested with restriction enzyme *Pst*I and hybridised to a probe for cDNA A3B02. The lower band observed in the patient indicates the presence of an RFLP (Restriction Fragment Length Polymorphism).

The first db(EST) search on the sequence of cDNA clone 43911 identified 4 overlapping clones. Therefore, the clone with the largest insert was the one selected for further analysis.

4.3.9 cDNA library screening

The collaboration with the Resource Centre of the German Human Genome Project identified one positive clone from a gridded cDNA library filter containing; liver, spleen, whole brain, skin, eye, ovary, lung, tonsil, melanocyte, pregnant uterus, heart, colon, prostate, kidney, thyroid, pancreas, and adrenal gland, following hybridisation with probe A3B02. The sequence generated from positive clone B1 overlapped with clone A3B02 with 100% homology.

Fourteen positive clones were identified with probe 43911. Direct sequencing of the five positive clones with the largest inserts showed them all to overlap with clone 43911 with 100% homology. However, none of the 5 clones extended the sequence of clone 43911 and were discarded from this point.

4.3.10 RACE PCR

Two RACE PCR products were generated from the human small intestine and human testis cDNA libraries with gene specific primers designed from the 5' end sequence of cDNA clone A3B02, see Figure 4.7. Direct sequencing of both products generated sequence that overlapped with clone A3B02 with 100% homology and extended the cDNA.

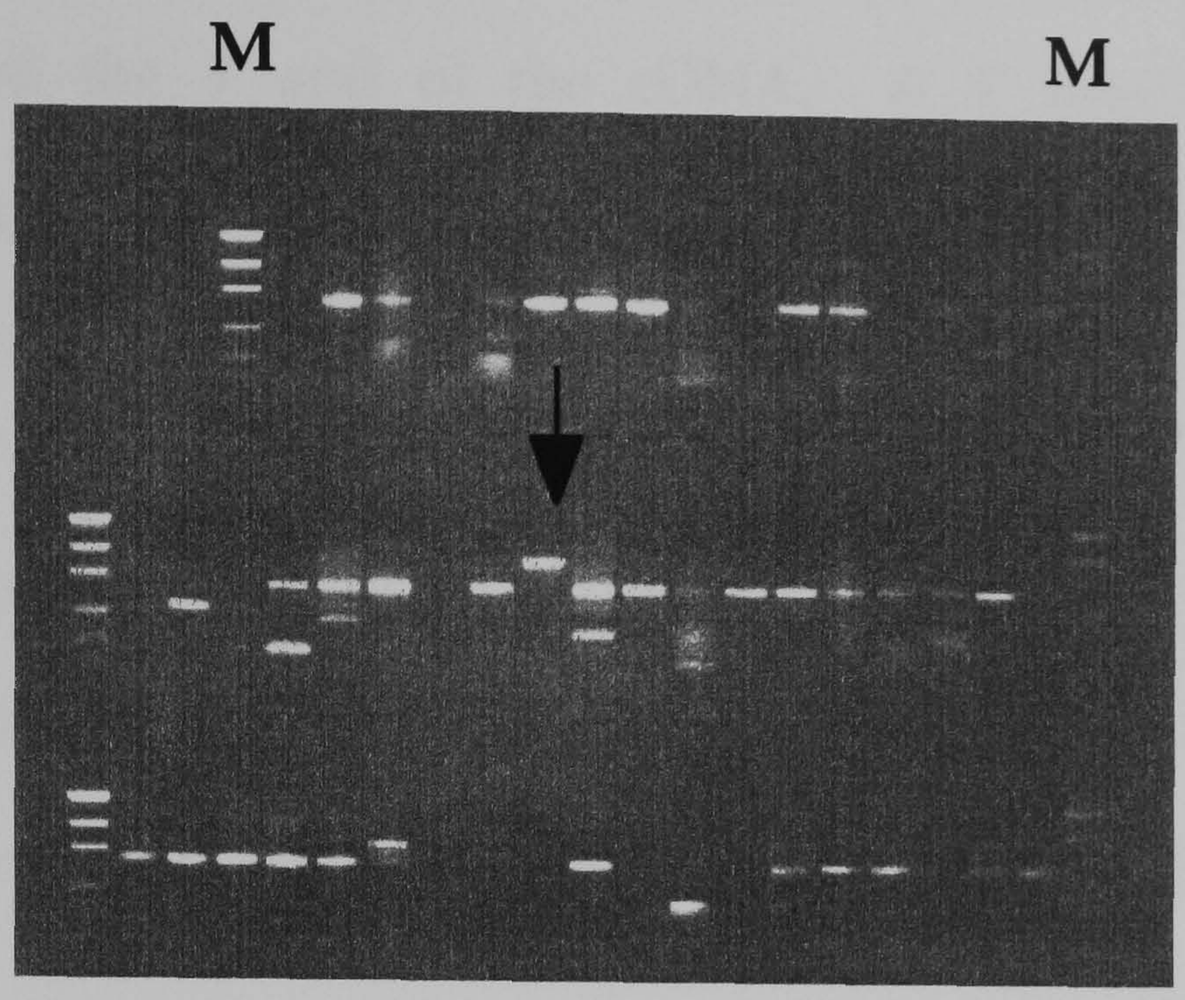


Figure 4.7

Representative 3' RACE PCR analysis of I.M.A.G.E. cDNA clone A3B02. Marathon-Ready™ cDNAs were primed with GSP A3B02R3 and AP1, then nested with GSP A3B02R1 and AP2. Products were analysed on a 1% agarose gel. The arrow indicates the 929bp product from the human small intestine cDNA library. The PCR products were sized with the Low Mass DNA marker (M).

Two RACE PCR products were generated from the human pituitary gland and human testis cDNA libraries with gene specific primers from the 3' end sequence of cDNA clone 43911. The RACE PCR products included a 13bp Poly-$(A)^+$ tail. This confirmed the 3' end of the cDNA. A 5' RACE PCR product was subsequently generated from the human foetal skeletal muscle cDNA library to extend the sequence at the 5' end.

Four 5' RACE PCR products were generated to complete the full coding sequence of novel cDNA H82831. The products were generated from the human pituitary gland, human testis, human placenta, and human skeletal muscle cDNA libraries.

4.3.11 Localisation to the YAC contig

All three novel cDNAs were shown to be sublocalised to YAC 816D6 encompassing the critical region of the 5q- syndrome at 5q31.3-q33.

4.3.12 Expression analysis

Subsequent RT-PCR analysis showed all three novel cDNAs to be amplified in RNA extracted from $CD34^+$ cells.

4.3.13 Database analysis using the Genetics Computer Group (GCG) software package

FastA nucleotide and BlastX protein homology searches utilising the GenEMBL and SWISS-PROT databases respectively, showed the novel cDNAs to have no known homology with any gene, identifying them as completely novel. The sequence generated from original cDNA clone A3B02 was submitted to GenBank,

and assigned the Accession number AF010242 (Boultwood *et al.*, 1997). The sequence generated from original cDNA clone 43911 was submitted to GenBank, and assigned the Accession number AF156165 (Boultwood *et al.*, 2000).

4.3.14 Novel cDNA AF010242

The 2410bp sequence of novel cDNA AF010242 was generated firstly from two cDNA clones and secondly from two RACE PCR products. A FastA nucleotide homology search showed the second RACE PCR product to have 100% homology with the human synaptopodin gene over 169bp. This result confirmed novel cDNA AF010242 to be the 3' untranslated region (UTR) of the human synaptopodin gene. The localisation of the human synaptopodin gene to chromosome 5 and the critical region of the 5q- syndrome has subsequently been mapped by others at the HGP using the Ensembl program.

4.3.15 Novel cDNA AF156165

The 2167bp sequence of novel cDNA AF156165 was generated from two cDNA clones and two RACE PCR products and submitted to GenBank as a *Homo sapiens* putative tumour suppressor mRNA. Until recently, the Human Gene Map and the UniGene set described cDNA AF156165 as a novel transcript. However, a subsequent database analysis showed novel cDNA AF156165 to be the 3' UTR of the human dynactin *p62* gene (Karki *et al.*, 2000). The localisation of dynactin *p62* to chromosome 5q has been confirmed by others at the HGP using the Ensembl program.

4.3.16 Novel cDNA H82831

The 1845 bp sequence of novel cDNA H82831 was generated from one cDNA clone and four RACE PCR products. Subsequently, H82831 was blasted against the Ensembl database. The 1845bp sequence of H82831 was found within contig AC034205, with 100% homology, proximal to the *MEGF1* gene at 5q32. Contig AC034205 contains 148198bp of 'working draft' sequence currently consisting of 13 ordered pieces (contigs). Novel gene H82831 lies within contig 11. However, Ensembl had not predicted H82831 as a novel gene. This may be due to the MER19 repeats in the sequence that may mask any motifs preventing it from being predicted. Alternatively, it may not be predicted because it is not a homologue, and does not have any protein motifs, i.e. is completely novel.

4.3.17 Summary

Twenty-three novel cDNAs were identified from the approximate 5Mb critical region of the 5q- syndrome at 5q31-q33 flanked by the genes *FGF1* and *IL12ß*. During the study a new patient with MDS and a 5q deletion was identified. This narrowed the critical region to approximately 3Mb at 5q31.3-q33 flanked by the genes *ADRß2* and *IL12ß*. Twenty-two percent of the cDNAs were then shown to map outside the new region, and excluded from the study.

The three novel cDNAs selected for further analysis fulfilled the following criteria: localisation to the critical region of the 5q- syndrome at 5q31.3-q33, expression in haematological tissues, and transcript sizes ≤4.4kb. 20/23 (87%) did not fulfil these criteria, and therefore not selected for further analysis.

4.4 Discussion

4.4.1 Mapping of chromosome 5-specific ESTs to the YAC contig

The collaboration with Professor Charles Auffray at Genethon was used as the primary resource to identify novel genes mapping to the approximate 5Mb critical region of the 5q- syndrome at 5q31-q33, flanked by the genes *FGF1* and *IL12ß*. Six chromosome 5-specific ESTs were PCR localised to the YAC contig spanning the critical region of the 5q- syndrome (Boultwood *et al.*, 1997).

4.4.1.1 Novel gene AF010242

Novel gene AF010242 was shown to map to the critical region of gene loss by gene dosage analysis. The cDNA was then sublocalised to YAC 816D6 from the YAC contig spanning the critical region of the 5q- syndrome (Li *et al.*, 1994; Kostrzewa *et al.*, 1998). Northern analysis showed AF010242 to possess a single transcript of 4.4kb and to be expressed in heart and skeletal muscle, and highly expressed in bone marrow, suggesting a possible role in leukaemogenesis. Further investigation showed novel gene AF010242 to have 100% homology with the 3' UTR of the human synaptopodin gene. Synaptopodin represents a novel, proline-rich, actin-associated protein that may play a role in modulating actin-based shape and mobility of dendritic spines and podocyte foot processes (Mundel *et al.*, 1997).

4.4.2 Novel gene identification from the Human Chromosome 5 GeneMap'98

In 1998, Jaju *et al.*, used FISH on the two 5q- syndrome patients that defined the CDR, along with a new, third patient with the 5q- syndrome and a small deletion, del(5)(q33q34), to refine further the critical deleted region. This resulted in the narrowing of the CDR within 5q31.3-5q33 to approximately 3Mb, flanked by the *ADRß2* and *IL12ß* genes.

Following the identification of patient 3, the Human chromosome 5 GeneMap'98 at NCBI was accessed to identify novel genes mapping to the 'new' approximate 3Mb critical region of the 5q- syndrome at 5q31.3-q33, flanked by the genes *ADRß2* and *IL12ß*. Five transcripts represented by EST sequences were identified.

4.4.2.1 Novel gene AF156165

Novel gene AF156165 was accessed from the GeneMap'98 at DNA marker D5S470. It was sublocalised to YAC 816D6 from the YAC contig by PCR screening. The 2167bp novel cDNA sequence was generated from two cDNA clones and two RACE PCR products. Novel gene AF156165 was submitted to GenBank as a *Homo sapiens* putative tumour suppressor mRNA. However, an updated UniGene search showed we had identified the 3' UTR of the human dynactin *p62* gene. Dynactin is a multisubunit complex and a required cofactor for most, or all, of the cellular processes powered by the microtubule-based motor cytoplasmic dynein (Karki *et al.*, 2000). The p62 subunit of dynactin was recently isolated using a dynein affinity column. Sequence analysis of the p62 polypeptide revealed a highly conserved N-terminal cysteine-rich domain, of which part of it fits a Zn^{2+}-binding RING domain (Karki *et al.*, 2000).

4.4.3 Novel gene identification from the Human Chromosome 5 GeneMap'99

Following the increase of deposited human sequences into db(EST), the GeneMap'99 was accessed between DNA markers D5S410 and D5S487 to identify novel genes mapping to the critical region of the 5q- syndrome. Twelve transcripts represented by EST sequences were identified.

4.4.3.1 Novel cDNAs AA040631, R76720, and H82404

Novel cDNAs AA040631 R76720 were found to have 100% homology to the previously identified human *CNOT8* (formerly *POP2*) gene. The *CNOT8* gene was previously cloned in this laboratory (Fidler *et al.*, 1999), and may play a role in the control of transcription.

Novel cDNA H82404 was found to have 100% homology to the human *H19* gene. The *H19* gene maps to 11p15.5 and is expressed in differentiating foetal cells (Hao *et al.*, 1993). This confirms the lack of signal in hybrid 5 and the very high level of expression seen in foetal liver in the Northern analysis. Studies have shown *H19* to be an imprinted gene with an important role in foetal differentiation, as well as a postulated function as a tumour suppressor gene (Doyle *et al.*, 1996). This result shows the *H19* gene to be incorrectly localised on the database and is unlikely to be involved in the pathogenesis of the 5q- syndrome.

4.4.4 Redundancy in the EST database

The results obtained in this study highlights the enormous value of the Human Chromosome 5 GeneMap, db(EST), and the UniGene set at NCBI but reveals the redundancy in the data. For example, the twelve transcripts identified from the Human GeneMap'99 in this study were assigned to DNA markers D5S410-D5S487 which flank the critical region of the 5q- syndrome at 5q31.3-q33. However, two of these (17%) were shown to be negative for chromosome 5 localisation by somatic cell hybrid analysis. One of these ESTs, representing an unidentified transcript, was shown to represent the human *H19* gene, previously localised to 11p15.5 (Hao *et al.*, 1993). Moreover, a further two ESTs, representing unidentified transcripts, were shown to represent the human *POP2* gene (Fidler *et al.*, 1999). The inaccuracy of the I.M.A.G.E. consortium distributors was also highlighted when screening db(EST) for overlapping cDNA clones. For example, three of nine

(33%) overlapping cDNA clones for novel gene *C5orf4* (Chapter 3) were found not to contain their EST sequence and therefore not the clone requested.

Others have described similar database errors. Bezieau *et al.*, (1998) showed that eleven of fifty (22%) of the ESTs assigned to the critically deleted region at 13q14.3 in B-CLL from the NCBI Human GeneMap and the Whitehead Institute Centre for Genome Research map, did not map to a YAC contig encompassing this critical region by PCR amplification. Similarly, nine of fifteen ESTs (60%) assigned to the 4Mb critical region of the hereditary paragangliomas at 11q23 by the Radiation Hybrid Mapping Consortium have been localised outside a YAC contig spanning this region (Baysal *et al.*, 1997).

Physical maps based on YAC clones (and especially CEPH mega-YACs) may be unreliable, however, because of high rates of chimaerism and deletions. This may explain the conflicting mapping data between the cDNA Radiation Hybrid Mapping Consortium and workers utilising YAC contigs at disease loci for EST mapping (Baysal *et al.*, 1997; Bezieau *et al.*, 1998; Liu *et al.*, 1998). Gene dosage analysis was used in this study to localise ESTs assigned between the DNA markers D5S410 and D5S487 prior to mapping within the YAC contig, thus reducing the likelihood of such errors. It is therefore important that independent mapping studies are performed to produce more refined and accurate EST-based transcription maps of a given genomic region. This is of particular importance at disease loci where ESTs often represent the first step in candidate gene isolation.

4.4.5 Human Genome Project progress

The rapid progression of the Human Genome Project has led to an increase in the number of ESTs deposited in db(EST) that represent novel coding sequences. The goal of the HGP to sequence the entire genome and identify the approximately 30,000 genes has led to the rapid characterisation of these novel cDNAs. For

example, two novel cDNAs (AF010242 and AF156165) from this study were found to represent the 3' UTRs of the synaptopodin and dynactin *p62* genes respectively. Therefore, the next stage of this study will utilise the data provided by the HGP and select candidate genes for mutation analysis with the aim of identifying the putative tumour suppressor gene associated with the development of the 5q- syndrome.

4.5 Conclusion

The collaboration with Professor Charles Auffray at Genethon, the Human Chromosome 5 GeneMaps'98 and '99, and the UniGene set has been successfully used to isolate nine novel coding sequences that map to the 5Mb critical region of the 5q- syndrome. All nine cDNAs were expressed in haematological tissues and represented candidates for the putative tumour suppressor gene associated with the development of the 5q- syndrome. Two of the novel cDNAs, AF010242 and AF156165 were identified as the human synaptopodin and dynactin *p62* genes respectively.

During the study, a new patient with MDS and the 5q- syndrome was identified. Cytogenetic analysis defined a small deletion with the proximal breakpoint at 5q31.3. This patient, therefore, narrowed the critical region from approximately 5Mb (flanked by the genes *FGF1* and *IL12ß*), at5q31-q33, to approximately 3Mb (flanked by the genes *ADRß2* and *IL12ß*) at 5q31.3-q33. The proximal breakpoint of the new critical region of the 5q- syndrome excluded novel cDNAs Cdy-17a06 and Bda-87b11.

This study has highlighted the advantages and disadvantages of the EST resource. The Human GeneMap at NCBI identified seventeen transcripts that represented candidates for the 5q- syndrome putative tumour suppressor gene. Conflicting mapping data excluded several of these transcripts while others were incomplete due to lack of 5' ESTs. Moreover, eight of seventeen (47%) ESTs represented large transcripts over 7.5kb. Transcript sizes less than 4.4kb were considered higher priority in the study.

There are many already known genes mapping to the approximate 3Mb critical region of the 5q- syndrome at 5q31.3-q33 with known/predicted functions that make them candidates for the 5q- syndrome putative tumour suppressor gene. For example, genes that regulate the cell cycle, have antioxidant properties and possess tumour suppressor activity. Therefore, mutation analysis on known genes mapping to the critical region of the 5q- syndrome was carried out alongside the isolation of novel coding sequences.

Chapter 5

Analysis of species homologous ESTs mapping to the critical region of the 5q- syndrome

5.1 Introduction
- 5.1.1 Comparative genomics
- 5.1.2 Identifying species homologous genes
- 5.1.3 Identifying species homologous genes in MDS and leukaemia
- 5.1.4 ESTs and the UniGene set
- 5.1.5 Isolating species homologous ESTs mapping to the critical region of the 5q- syndrome
  - 5.1.5.1 The Leucyl-tRNA synthetase, cytoplasmic (*CDC60*) gene
  - 5.1.5.2 The Regulator of mitotic spindle assembly 1 (*RMSA-1*) gene
  - 5.1.5.3 The Goliath protein
  - 5.1.5.4 The Protein phosphatase 2A beta subunit (*PP2A*) gene and the Protein phosphatase 1, regulatory (inhibitor) subunit 2 (*PPP1R2*) gene
  - 5.1.5.5 The Tetratricopeptide repeat protein (tpr1)
- 5.1.6 Aims of the study

5.2 Materials and methods
- 5.2.1 EST identification
- 5.2.2 I.M.A.G.E. cDNA clones
- 5.2.3 Samples
- 5.2.4 Gene dosage analysis
- 5.2.5 Northern analysis
- 5.2.6 Direct sequencing
- 5.2.7 Overlapping cDNA clones
- 5.2.8 cDNA library screening
- 5.2.9 Database analysis using the Genetics Computer Group (GCG) software package

5.3 Results
5.3.1 ESTs identified from the Human GeneMap and the UniGene set
5.3.2 Gene dosage analysis
5.3.3 Northern analysis
5.3.4 Direct sequencing
5.3.5 Overlapping cDNA clones
5.3.6 cDNA library screening
5.3.7 Summary
5.4 Discussion
5.5 Conclusion

5.1 Introduction

5.1.1 Comparative genomics

The completion of the sequencing of the *Saccharomyces cerevisiae* and *Caenorhabditis elegans* genomes has aided the understanding of sequence data currently being generated by the Human Genome Project. Comparative genomics – the cross-referencing of information between species – has been used to determine how the function and position of genes has changed over the course of evolution (Graves, 1998). *Fugu rubripes* (the Japanese puffer fish) is particularly suited to this kind of analysis because whilst its 400Mb genome is eight times smaller than human's, it has a similar repertoire of genes. These genome characteristics along with the large evolutionary distance between bony fish and mammals make *Fugu* a useful tool for studying gene evolution (Elgar *et al.*, 1999). Moreover, if regions of the two genomes exhibited conservation of gene order (i.e., were syntenic), it should be possible to dramatically reduce the effort required for identification of candidate genes in human disease loci by sequencing syntenic regions of the compact *Fugu* genome. For example, Trower *et al.*, (1996) demonstrated three genes (dihydrolipoamide succinyltransferase, S31iii125, and S20i15), which are linked to FOS in the familial Alzheimer disease locus (AD3) on human chromosome 14, to have homologues in the *Fugu* genome adjacent to *Fugu* cFOS. The relative gene order of cFOS, S31iii125, and S20i15 was the same in both genomes, but in *Fugu* these three genes lay within a 12.4kb region, compared to >600kb in the human AD3 locus. These results demonstrate the conservation of synteny between the genomes of *Fugu* and man and highlight the utility of this approach for sequence-based identification of genes in human disease loci.

Drosophila melanogaster has also been used as a model for comparative genomics. Rubin *et al.*, (2000) examined 289 human disease genes and found the fruitfly to have homology to 177 of them providing the foundation for rapid analysis of some of the basic processes involved in human disease. Comparisons over vast

evolutionary time scales show that the mammalian genome has been highly conserved. Thus, information about location and function of genes is directly transferable across species and should greatly accelerate the search for genes that specify inherited human diseases (Graves, 1998).

5.1.2 Identifying species homologous genes

The introduction of molecular biology techniques has allowed the isolation and identification of several oncogenes and tumour suppressor genes. Analysis of genetic alterations in these genes has enabled the comparison of carcinogenesis pathways in humans and rodents at the molecular level (Goodrow, 1996). The results from this study showed that most of the oncogenes/tumour suppressor genes found to be altered in humans were also altered in rodents. There are still many unknown steps in the process of carcinogenesis. However, overall, the results indicate that despite the differences between rodents and humans, the use and comparison of rodent models with human tumorigenesis is one of the best ways to examine the mechanisms of carcinogenesis (Goodrow, 1996).

D. melanogaster has been the target of extensive genetic analyses over the past ninety years and a notable amount of information is known about its gene structure, gene regulation and gene function. Banfi *et al.*, (1997) utilised the EST resource to identify novel human and murine gene transcripts homologous to *Drosophila* mutant genes. Mapping and expression studies were carried out in order to characterise these novel genes. The authors state that the comparison between these novel genes and their putative partners in *Drosophila* contributes to the understanding of their function in mammals and to the discovery of their possible role in disease.

5.1.3 Identifying species homologous genes in MDS and leukaemia

A number of genes involved in the pathogenesis of MDS and leukaemia have been shown to have homology to genes from other species. The *AML1* gene which is rearranged by the t(8;21) translocation in AML is highly homologous to the *Drosophila* segmentation gene and the mouse transcription factor PEBP2 alpha subunit gene (Miyoshi *et al.*, 1995). This region of homology, called the Runt domain, is responsible for DNA-binding and protein-protein interactions. Three mouse genomic domains, Fim1, Fim2, and Fim3 were previously described as proviral integration regions frequently involved in the early stages of myeloblastic leukaemogenesis induced *in vivo* or *in vitro* by the Friend murine leukaemia virus (Van Cong *et al.*, 1989). The human homologues of these three mouse domains, were found to correspond to human loci involved in genetic alterations specific to some human leukaemias. Fim2 was identified as the 5' end of the *c-FMS* protooncogene, which encodes the receptor of the macrophage colony stimulating factor. The functions of Fim1 and Fim3 are not yet known, but these regions are highly conserved among different species. Mapping of these human homologues showed that the localisation of *FIM2/c-FMS* on 5q was confirmed, while *FIM1* and *FIM3* were localised on human chromosomes 6p22.33-p23 and 3q27 respectively. Translocations involving these two regions have been described in various haematopoietic malignancies: the t(6;9) (p23;q34) in acute nonlymphocytic leukaemias and the 3q26-q28 translocations in a large variety of leukaemias (Van Cong *et al.*, 1989).

5.1.4 ESTs and the UniGene set

The EST database is now used by many researchers as the primary resource for identifying genes localised to a specific chromosomal region. UniGene (http://www.ncbi.nlm.nih.gov/UniGene/) is an experimental system for automatically partitioning GenBank EST sequences into a non-redundant set of

gene-oriented clusters. Each UniGene cluster contains sequences that represent a unique gene, as well as related information such as the tissue types in which the gene is expressed and map location (UniGene homepage). In addition to sequences of well-characterised genes, hundreds of thousands of novel EST sequences have been included. Human UniGene is updated every week with new EST sequences, and bimonthly with new characterised sequences. Currently, sequences from the human genome, rat, mouse, cow, zebrafish and clawed frog have been processed. These species were chosen because they have the greatest amounts of EST data available and represent a variety of species.

5.1.5 Isolating species homologous ESTs mapping to the critical region of the 5q- syndrome

There are three distinctions of similarity that UniGene clusters are assigned to: "Highly similar to" means >90% homology in the aligned region; "Moderately similar to" means 70-90% homology in the aligned region; and "Weakly similar to" means <70% homology in the aligned region, with other species. We selected ESTs representing transcripts that were highly similar to genes from other species, for further analysis. In addition to these ESTs representing human homologues, we selected ESTs representing transcripts that were the complete coding sequences of human genes (i.e., human mRNAs). The human genes selected for analysis have been implicated in the aetiology of tumours, and therefore represent candidates for the putative 5q- syndrome tumour suppressor gene. The ESTs selected for further analysis had similarity to the known proteins after translation and/or the corresponding clone source was a CGAP library.

Since the advent of this study, a further resource to identify human genes homologous to other species has become available. HomoloGene is a homology resource which includes both curated and calculated homologues for genes represented in UniGene and LocusLink for human, mouse, rat, cow, zebrafish,

frog and fly. The calculated homologues are the result of nucleotide sequence comparisons between all UniGene clusters for each pair of organisms. These homologues are considered putative since they are based only on sequence comparisons. Nucleotide sequences for each pair of organisms are compared to identify sequences pairs that share the highest degree of nucleotide sequence similarity. The best match for a sequence in one organism to a sequence in a second organism is based on the percent of identical sequence (%ID) in an alignment over a minimum 100 base pairs.

The EST database and UniGene have been used to identify ESTs mapping to the critical region of the 5q- syndrome with homology to known human genes or homology to genes from other species. In total, six transcripts were included in this study (three of which represented human homologues of yeast and *Drosophila* genes, and three which represented known human genes) each represented by at least two ESTs from db(EST) and the UniGene set. These transcripts represent candidates for the putative tumour suppressor gene(s) associated with the 5q- syndrome.

5.1.5.1 The Leucyl-tRNA synthetase, cytoplasmic (*CDC60*) gene

The yeast cell division cycle gene (*CDC60*) is indirectly involved in the regulation of the cell cycle (Hohmann and Thevelein, 1992). There are several tumour suppressor genes known to be involved in the control of the cell cycle that include *p16*, *p53*, and *RB1*.

5.1.5.2 The Regulator of mitotic spindle assembly 1 (*RMSA-1*) gene

The *RMSA-1* gene is essential for mitotic spindle assembly. The assembly of a bipolar spindle is essential for the accurate segregation of replicated chromosomes during cell division (Waters and Salmon, 1995). The *p53* tumour suppressor gene

is believed to be involved in the mitotic spindle checkpoint and in the regulation of centrosome function (Morgan and Kastan, 1997).

5.1.5.3 The Goliath protein

The Goliath protein (g1) is involved in the regulation of gene expression during mesoderm formation in *Drosophila*. It is also thought to have a putative role as a transcription factor (UniGene Hs.9788 data). The nucleotide sequence of its cDNA encodes a 32-kDa protein with two putative zinc fingers, and a serine/glutamine/proline-rich region. These features indicate a functional role for g1. Tumour suppressor genes encoding zinc fingers include the Wilms' tumour suppressor gene (*WT1*) and the Human Kruppel-related 3 (*HKR3*) gene. *WT1* encodes a zinc finger transcription factor that regulates expression of several genes involved in cellular proliferation and differentiation (Bardeesy and Pelletier, 1998). The *HKR3* gene maps within chromosome subbands 1p36.2-36.3, a region postulated to contain a tumour suppressor gene associated with advanced neuroblastomas (Maris *et al.*, 1997).

5.1.5.4 The Protein phosphatase 2A beta subunit (*PP2A*) gene and the Protein phosphatase 1, regulatory (inhibitor) subunit 2 (*PPP1R2*) gene

Phosphatases are regulatory enzymes that antagonise the action of kinases within the cell (Parsons, 1998). An understanding of the contribution of kinases to cancer has emerged during the past two decades. Currently, three phosphatases have been implicated in the aetiology of tumours: protein phosphatase 2A (*PP2A*), *CDC25A/B*, and *PTEN* (or *MMAC1*). *PP2A* and *PTEN* have been shown to function as tumour suppressor genes (Parsons, 1998).

5.1.5.5 The Tetratricopeptide repeat protein (tpr1)

The tetratricopeptide repeat (TPR) motif is a protein-protein interaction module found in multiple copies in a number of functionally different proteins that facilitates specific interactions with a partner protein(s). Most TPR-containing proteins are associated with multiprotein complexes, and there is extensive evidence indicating that TPR motifs are important to the functioning of chaperone, cell-cycle, transcription, and protein transport complexes (Blatch and Lassle, 1999). Tumour suppressor genes that harbour this repeat have been identified. Loss of heterozygosity in 1p31 is a frequent genetic alteration in breast tumours indicating the site of a tumour suppressor gene (Su *et al.*, 1999). Su *et al.*, isolated a new member of the human tetratricopeptide repeat-containing family of genes, *TTC4*, which maps to this region. Other members of this gene family have been implicated in tumourigenesis suggesting that *TTC4* may represent a breast cancer tumour suppressor gene (Su *et al.*, 1999).

5.1.6 Aims of the study

The aims of this study were to use the EST resource to identify human homologues of genes previously identified in other species, and to map them to the transcript map currently being generated for the critical region of gene loss in the 5q- syndrome. These known genes would be localised in relation to the novel coding sequences previously assigned to the transcript map (Chapter 4).

Over the last ten years, several human genes have been cloned based on their homology to genes previously identified in model organisms. For example, Bronner *et al.*, (1994) proposed that the *hMLH1* (human MutL homologue) was the HNPCC (hereditary non-polyposis colon cancer) gene located on 3p because of the

similarity of the *hMLH1* gene product to the yeast DNA mismatch repair protein, MLH1. In relation to this study, Tugendreich *et al.*, (1993) used the EST database to identify and positionally map human homologues of yeast genes to cross-reference the biological and genetic information known about yeast genes to mammalian chromosomal maps. The authors scanned db(EST) for human open reading frames related to yeast protein sequences and used the corresponding human cDNA to obtain a high-resolution map position on human and mouse chromosomes. Their results identified the human homologue of *S. cerevisiae* CDC27 that mapped to human chromosome 17 and mouse chromosome 11.

The identification and mapping of human homologues to the critical region of the 5q- syndrome could facilitate the identification of the 5q- syndrome tumour suppressor gene.

5.2 Materials and Methods

5.2.1 EST identification

The Human GeneMap and the UniGene set at NCBI were accessed for species homologous ESTs and ESTs representing known human genes, mapping to the YAC contig spanning the approximate 5Mb critical region of the 5q- syndrome at 5q31-q33, flanked by the genes *FGF1* and *IL12ß*. The search revealed six transcripts. Three transcripts were represented by ESTs 'highly similar' to known genes from other organisms (e.g. *Drosophila melanogaster*, and *Saccharomyces cerevisiae*). The remaining three transcripts were represented by ESTs from the complete cds of a known human gene. Two ESTs representing each transcript were selected for further analysis. The ESTs selected had similarity to known proteins (after translation), contained a polyadenylation signal, contained a mapped sequence-tagged site (STS), and its clone source was a CGAP library.

5.2.2 I.M.A.G.E. cDNA clones

I.M.A.G.E. cDNA clones of the ESTs were obtained from the UK HGMP Resource Centre at Hinxton, Cambridge as stabs in agar. Single colonies were obtained by plating onto LB ampicillin (50mg/ml) plates. A single colony was then inoculated into a 10ml LB culture containing ampicillin. Plasmid DNA was obtained using the QIAprep® Spin Miniprep Kit. The insert was excised using the appropriate restriction enzymes and purified for use as a probe with the Wizard® PCR Preps DNA Purification System.

5.2.3 Samples

Four patients with the classical features of the 5q- syndrome, including the two patients that defined the approximate 5Mb critical region of the 5q- syndrome at 5q31-q33, were included in the study. Granulocyte and mononuclear cells were

separated from 40mls of peripheral blood by ficoll gradient centrifugation (Boyum, 1984). The granulocytes showed a high level of purity (≥95%). Mononuclear cells (specifically T-lymphocytes) were isolated by erythrocyte rosetting and showed a purity of ≥90%. High molecular weight DNA was obtained from the fractionated blood leukocytes by Nucleon® extraction. Granulocyte DNA fractions from the peripheral blood of healthy individuals were used as controls. High molecular weight DNA was obtained from a human/mouse hybrid cell line with human chromosome 5 as its only human complement.

5.2.4 Gene dosage analysis

Gene dosage analysis was used to confirm the localisation of the EST (represented by its I.M.A.G.E. cDNA clone) to human chromosome 5; and to determine the loss or retention of the gene in the patient granulocyte DNA (**Chapter 3 section 3.2.4**). Gene dosage experiments were carried out on at least two separate occasions.

5.2.5 Northern analysis

I.M.A.G.E. cDNA clones which hybridised to a single fragment in hybrid 5 DNA and showed a 50% dosage reduction were hybridised to Multiple Tissue Northern (MTN) blots (**Chapter 3 section 3.2.5 and Table 3.1**).

5.2.6 Direct sequencing

I.M.A.G.E. cDNA clones representing each EST were sequenced as either single-stranded or double-stranded templates, as previously described, by the dideoxy chain termination method (Sanger *et al.*, 1977) (**Chapter 2 section 2.12.1**). Clones were sequenced using the Cy5 Autoread sequencing kit (**Chapter 3 section 3.2.7**). Each I.M.A.G.E. cDNA clone was sequenced in full and then subjected to a

GenBank homology search to confirm its homology with the known gene, and to identify overlapping clones to generate the full-length cDNA.

5.2.7 Overlapping cDNA clones

Sequence data from each I.M.A.G.E. cDNA clone was subjected to a homology search against the EST database db(EST) at NCBI for overlapping cDNA clones to generate the full-length cDNA (**Chapter 3 section 3.2.8**). Sequence data from the overlapping clone was added to the sequence from the I.M.A.G.E. cDNA clone, and the 'new' sequence submitted to db(EST).

5.2.8 cDNA library screening

If no overlapping clones were identified from db(EST) or UniGene, the cDNA clone insert was screened against cDNA libraries. In the first instance, a foetal brain cDNA library was selected as this tissue expresses a wide variety of genes. Seven high-density gridded cDNA filters were used in the study. Also, a collaboration with the Resource Centre of the German Human Genome Project at the Max-Planck-Institute for molecular genetics (RZPD) was established (**Chapter 3 section 3.2.9**). Positive clones were sequenced and the 'new' sequence submitted to db(EST).

5.2.9 Database analysis using the Genetics Computer Group (GCG) software package

FastA, BlastX and Frames analysis was carried out on the sequence generated from the I.M.A.G.E. cDNA clones, as previously described (**Chapter 3 sections 3.2.12.1-3.2.12.3**).

5.3 Results

5.3.1 ESTs identified from the Human GeneMap and the UniGene set

ESTs identified from the Human Chromosome 5 GeneMap at NCBI, and the UniGene set are displayed in Table 5.1. I.M.A.G.E. cDNA clones from which each EST was originally derived were obtained.

5.3.2 Gene dosage analysis

Three out of six (50%) I.M.A.G.E. cDNA clones representing ESTs 'highly similar' to *CDC60*, *RMSA-1*, and *PP2A* were shown to map to the 5Mb critical region of the 5q- syndrome at 5q31-q33. An approximate 50% reduction in the dosage of each cDNA clone in the granulocyte patient DNA compared with normal controls, confirmed the deletion of one allele.

Gene dosage analysis with I.M.A.G.E. cDNA clone 308419 derived from EST W44992 (Goliath protein) showed that the 308419 probe hybridised to 18 fragments in the granulocyte DNA from the patients and controls, and 8 fragments in the Hybrid 5 DNA. Not one of the 18 fragments appeared to show a 50% reduction in the granulocyte patient DNA, suggesting cDNA clone 308419 did not map to the critical region of the 5q- syndrome as had been predicted by the Human GeneMap and the UniGene set. Moreover, mapping information from UniGene showed the gene had been mapped between two sets of DNA markers from the transcript map. No further analysis was carried out on EST W44992.

Table 5.1 ESTs identified from the Human Chromosome 5 GeneMap and UniGene set

I.M.A.G.E. clone name	GenBank Accession No.	D marker interval	Tissue source (cDNA library)	Species homology
33583	R44866	D5S402-D5S2090	Soares infant brain 1NIB	Highly similar to Leucyl-tRNA synthetase (*CDC60*) gene [*Saccharomyces cerevisiae*]
145513	R77718	D5S436-D5S470	Soares placenta Nb2HP	Highly similar to regulator of mitotic spindle assembly 1 (*RMSA-1*) gene [*Homo sapiens*]
308419	W44992	D5S2119-D5S402 D5S658-D5S402	Soares foetal lung NbHL19W	Highly similar to Goliath protein [*Drosophila melanogaster*]
145699	R78295	D5S410-D5S487	Soares placenta Nb2HP	Protein phosphatase 1, regulatory (inhibitor) subunit 2 (*PPP1R2*) gene [*Homo sapiens*]
40699	R55800	D5S436-D5S470	Soars infant brain 1NIB	Protein phosphatase 2A beta subunit (*PP2A*) gene [*Homo sapiens*]
194016	H51264	D5S412-D5S422	Soares foetal liver spleen 1NFLS	Tetratricopeptide repeat protein (tpr1) [*Homo sapiens*]

ESTs H51264 and R78295 representing the *tpr1* and *PPP1R2* genes respectively were shown to map outside the distal breakpoint of the critical region by gene dosage analysis, see Figure 5.1. A UniGene search later identified *PPP1R2* to map to chromosome 3q29.

Thus the *CDC60*, *RMSA-1*, and *PP2A* genes were selected for further analysis.

5.3.3 Northern analysis

Northern analysis showed I.M.A.G.E. clone 33583 (*CDC60*) to be ubiquitously expressed and to possess a transcript of 4.4kb in addition to a faint transcript of 6.0kb, see Figure 5.2.

5.3.4 Direct sequencing

Direct sequencing of clone 33583 (*CDC60*) generated 1073bp of sequence. A GenBank homology search showed the cDNA to be the human homologue of the yeast cell division cycle gene *CDC60* (leucyl-tRNA synthetase, cytoplasmic). A BlastX protein homology search utilising the SWISS-PROT database showed the cDNA to have a 46% amino acid match over its entire length with *Saccharomyces cerevisiae* CDC60 protein and a 43% amino acid match with *Neurospora crassa* leucyl-tRNA synthetase, cytoplasmic, see Figure 5.3.

Direct sequencing of I.M.A.G.E. clone 145513 derived from EST R77718 (*RMSA-1*) generated 561bp of sequence. A FastA nucleotide homology search showed clone 145513 to have 97.1% identity with a *Homo sapiens* chromosomal protein, see Figure 5.4.

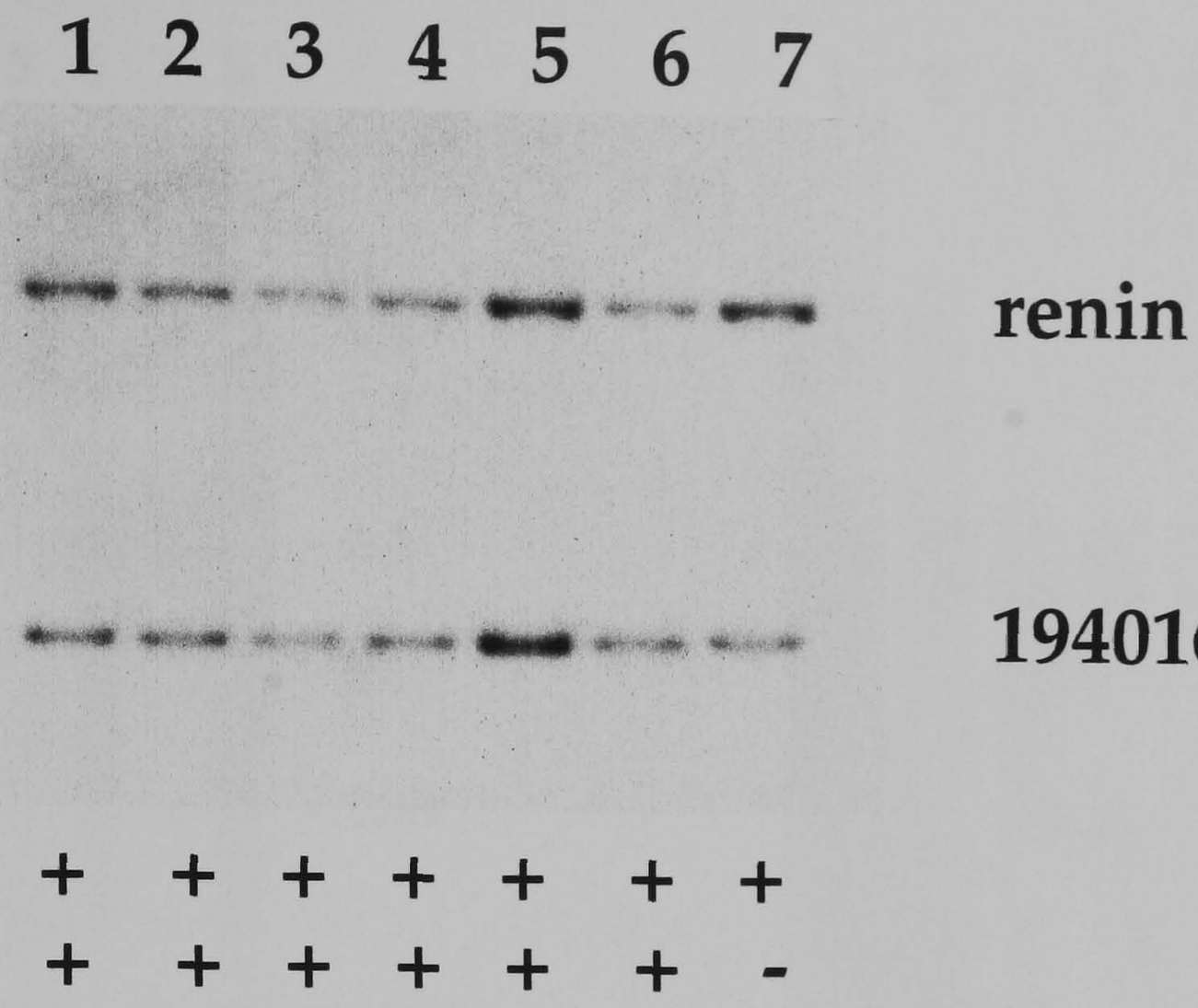


Figure 5.1

Representative gene dosage analysis of I.M.A.G.E. cDNA clone 194016 from EST H51264 (*tpr1*). DNA obtained from the granulocyte fractions of 3 patients (lanes **1**, **3**, and **7**), the lymphocyte fractions from 2 patients (lanes **2** and **4**), and healthy controls (lanes **5** and **6**) was digested with *Eco*RI and simultaneously hybridised to a probe for 194016 and a probe for the *renin* gene. ++ indicates the presence of two copies of the *tpr1* gene and + - indicates the deletion of one copy of the gene.

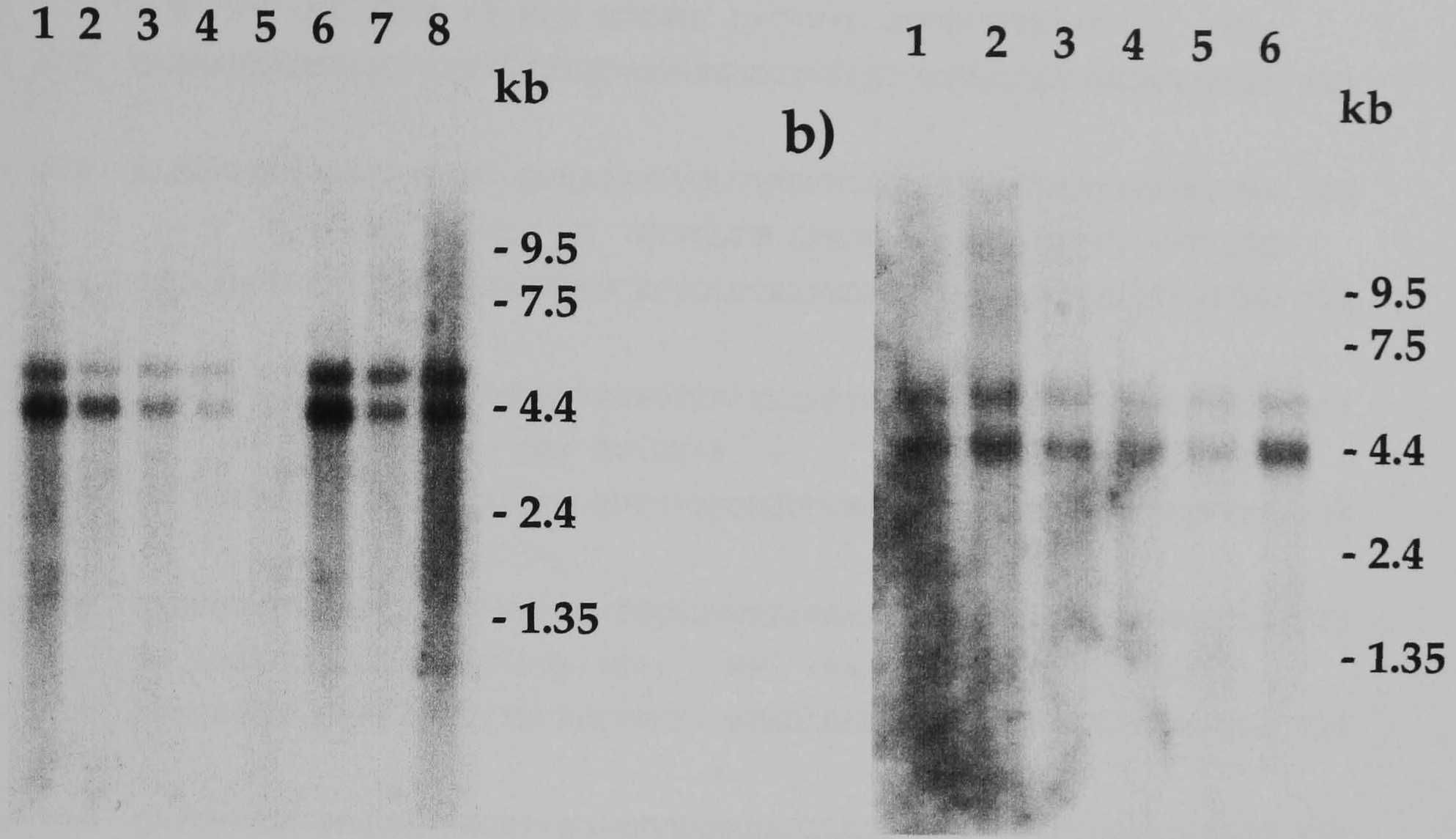


Figure 5.2

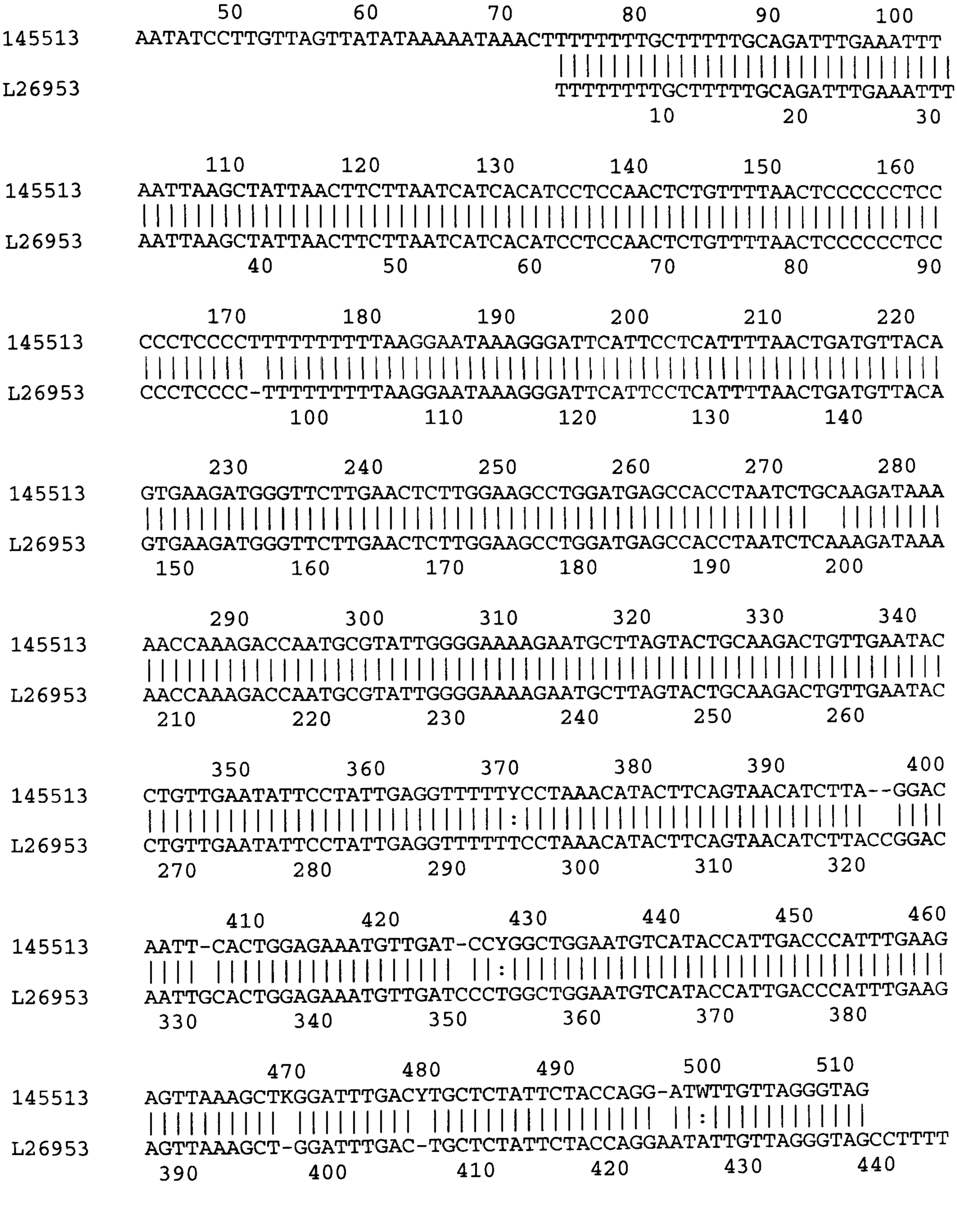
Representative Northern blot analysis of I.M.A.G.E. cDNA clone 33583 (*CDC60*). The MTN blot (a) included 2µg of poly (A^+) RNA from; heart (1), brain (2), placenta (3), lung (4), liver (5), skeletal muscle (6), kidney (7), and pancreas (8). The MTN blot (b) included 2µg of poly (A^+) RNA from; spleen (1), lymph node (2), thymus (3), peripheral blood leukocytes (4), bone marrow (5), and foetal liver (6). Sizes of RNA marker bands (kb) are indicated approximately.

Figure 5.3 **BlastX analysis of I.M.A.G.E. cDNA clone 33583 (top) and the yeast cell division cycle gene *CDC60* (bottom)**

```
33583:  6    EVKKTIQKKMIDAGDALIYMEPEKQVMSRSSDECVVALCDQWYLDYGEENWKKQTSQCLK 185
             + K  ++  MI AG+A +Y EPE QVMSRS D+C+V+L DQWY+DYGEE+WKKQ  +CL+
CDC60: 505   QAKNKVKADMIAAGEAFVYNEPESQVMSRSGDDCIVSLEDQWYVDYGEESWKKQAIECLE 564

33583: 186   NLETFCEETRRNFEATLGWLQEHACSRTYGLGTHLPWDEQWLIESLSDSTIYMAFYTVAH 365
              ++ F  E +  FE  L WL+  A  RTYGLGT LPWDE++L+ESLSDSTIY +FYT+AH
CDC60: 565   GMQLFAPEVKNAFEGVLDWLKNWAVCRTYGLGTRLPWDEKYLVESLSDSTIYQSFYTIAH 624

33583: 366   LLQGGNLHGQAESPLGIRPQQMTKEVWDYVFFKEAPFPKTQIAKEKLDQLKQEFEFWYPV 545
              LL   + +G    PLGI   QMT EV+DY+F  +     T I    L +L++EFE++YP+
CDC60: 625   LL-FKDYYGNEIGPLGISADQMTDEVFDYIFQHQDDVKNTNIPLPALQKLRREFEYFYPL 683

33583: 546   DLRVSGKDLVPNHLSYYLYNHVAMWPEQSDKWPTAVRANGHLLLNSEKMSKSTGNFLTLT 725
              D+ +SGKDL+PNHL++++Y HVA++P++   WP  +RANGHL+LN+ KMSKSTGNF+TL
CDC60: 684   DVSISGKDLIPNHLTFFIYTHVALFPKKF--WPKGIRANGHLMLNNSKMSKSTGNFMTLE 741

33583: 726   QAIDKFSADGMRLALADAGDTVEDANFVEAMADAGILRLYTWVEWVKEMVANWDSLRSGP 905
              Q ++KF AD  R+A ADAGDTVEDANF E+ A+A ILRL+   EW +E +    +LR+G
CDC60: 742   QTVEKFGADAARIAFADAGDTVEDANFDESNANAAILRLFNLKEWAEE-ITKESNLRTGE 800

33583: 906   ASTFNDRVFASELNAGIIKTDQNYEKMMFKEALKTGFFEFQAAKDKYRELAVEGMHRELV 1085
               + F D  F  E+NA I KT + Y    +K ALK G F+FQAA+D YRE A   MH++L+
CDC60: 801   ITDFFDIAFEHEMNALIEKTYEQYALTNYKNALKYGLFDFQAARDYYRE-ASGVMHKDLI 859

33583: 1086  FRFIEVQTLLLAPFCPHLCEHIW-TLLGKPDSIMNASWPVAG-PVNEVLIHSSQYLMEVT 1259
              R+IE Q LLLAP  PH  E+I+  +LG   S+ NA +P A  PV++ ++ +  YL  +
CDC60: 860   ARYIETQALLLAPIAPHFAEYIYREVLGNQTSVQNAKFPRASKPVDKGVLAALDYLRNLQ 919

33583: 1260  HDLRLRLKNYMMPAKGKKTDKQPLQKPSHCTIYVAKNYPPWQHTTLSVLRKHFEANNGKL 1439
              +R      +   KGK  +     KP   T+ +++++P WQ   + ++RK F      L
CDC60: 920   RSIREGEGQALKKKKGKSAEID-ASKPVKLTLLISESFPEWQSQCVEIVRKLFSEQT--L 976

33583: 1440  PDNKVIASELGSMPELKKYMKKVMPFVAMIKENLEKMGPR-ILDLQLEFDE 1589
              DNK +   +       K MK+ MPF++++K+ L    P  + + +L+F E
CDC60: 977   DDNKKVREHIE-----PKEMKRAMPFISLLKQRLANEKPEDVFERELQFSE 1022
```

I.M.A.G.E. cDNA clone 33583 has a 46% amino acid match with the *Saccharomyces cerevisiae* CDC60 protein in a 1090bp overlap.

Figure 5.4 **Alignment of I.M.A.G.E. cDNA clone 145513 (top) and the *Homo sapiens* chromosomal protein mRNA (the human homologue of the *Drosophila RMSA-1* gene – GenBank accession No. L26953) (bottom)**



I.M.A.G.E. cDNA clone 145513 has a 97.1% identity with the *Homo sapiens* chromosomal protein mRNA over a 444bp nucleotide overlap.

This result suggests EST R77718 represents the human homologue of the *Drosophila RMSA-1* gene. A BlastX protein homology showed clone 145513 to have an 82%-97% amino acid match over two pieces of sequence with the *Drosophila* regulator of mitotic spindle assembly 1 (*RMSA-1*) gene.

5.3.5 Overlapping cDNA clones

db(EST) searches using the sequence of clone 33583 (*CDC60*) identified nine overlapping clones. Two of these I.M.A.G.E. clones, 567178 and 1173393 possessed large inserts, and were thus selected for further analysis. Direct sequencing of these overlapping clones generated sequence that overlapped with *CDC60* with 100% homology over part of the sequence. Both clones also generated a further 1218bp of sequence. A FastA nucleotide homology search on the new 2291bp sequence showed clone 33583 to have 99.7% identity with the *Homo sapiens* mRNA for leucyl transferase, see Figure 5.5. This result suggests EST R44866 represents the human homologue of the yeast *CDC60* gene.

5.3.6 cDNA library screening

The collaboration with the Resource Centre of the German Human Genome Project identified seven positive clones from the Soares infant brain and human foetal brain cDNA libraries following hybridisation with probe 33583 (*CDC60*). The sequence generated from 3/7 (43%) positive clones overlapped with clone 33583 with 100% homology but did not add any additional sequence. The seven clones were discarded from the study at this point.

Figure 5.5 **Alignment of I.M.A.G.E. cDNA clone 33583 (top) and the *Homo sapiens* mRNA for leucyl tRNA synthetase (human homologue of the yeast *CDC60* gene - GenBank accession No. D84223) (bottom)**

```
                                                    10        20        30
33583                                               GGCACGAGGTAAAGAAGACTATTCAGAAAAAG
                                                    || || ||||||||||||||||||||||||||
D84223          TTGGTGGATGGATTTAAAGGACAGAAGGTTCAAGATGTAAAGAAGACTATTCAGAAAAAG
                    1510      1520      1530      1540      1550      1560

                         40        50        60        70        80        90
33583           ATGATTGACGCTGGAGATGCACTTATTTACATGGAACCAGAGAAACAAGTGATGTCCAGG
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
D84223          ATGATTGACGCTGGAGATGCACTTATTTACATGGAACCAGAGAAACAAGTGATGTCCAGG
                    1570      1580      1590      1600      1610      1620

                        100       110       120       130       140       150
33583           TCGTCAGATGAATGTGTTGTGGCTCTGTGTGACCAGTGGTACTTGGATTATGGAGAAGAG
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
D84223          TCGTCAGATGAATGTGTTGTGGCTCTGTGTGACCAGTGGTACTTGGATTATGGAGAAGAG
                    1630      1640      1650      1660      1670      1680

                        160       170       180       190       200       210
33583           AATTGGAAGAAACAGACATCTCAGTGCTTGAAGAACCTGGAAACATTCTGTGAGGAGACC
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
D84223          AATTGGAAGAAACAGACATCTCAGTGCTTGAAGAACCTGGAAACATTCTGTGAGGAGACC
                    1690      1700      1710      1720      1730      1740

                        220       230       240       250       260       270
33583           AGGAGGAATTTTGAAGCCACCTTAGGTTGGCTACAAGAACATGCTTGCTCAAGAACTTAT
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
D84223          AGGAGGAATTTTGAAGCCACCTTAGGTTGGCTACAAGAACATGCTTGCTCAAGAACTTAT
                    1750      1760      1770      1780      1790      1800

                        280       290       300       310       320       330
33583           GGTCTAGGCACTCACCTGCCTTGGGATGAGCAGTGGCTGATTGAATCACTTTCTGACTCC
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
D84223          GGTCTAGGCACTCACCTGCCTTGGGATGAGCAGTGGCTGATTGAATCACTTTCTGACTCC
                    1810      1820      1830      1840      1850      1860

                        340       350       360       370       380       390
33583           ACTATTTACATGGCATTTTACACAGTTGCACACCTATTGCAGGGGGGTAACTTGCATGGA
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
D84223          ACTATTTACATGGCATTTTACACAGTTGCACACCTATTGCAGGGGGGTAACTTGCATGGA
                    1870      1880      1890      1900      1910      1920

                        400       410       420       430       440       450
33583           CAGGCAGAGTCTCCGCTGGGCATTAGACCGCAACAGATGACCAAGGAAGTTTGGGATTAT
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
D84223          CAGGCAGAGTCTCCGCTGGGCATTAGACCGCAACAGATGACCAAGGAAGTTTGGGATTAT
                    1930      1940      1950      1960      1970      1980

                        460       470       480       490       500       510
33583           GTTTTCTTCAAGGAGGCTCCATTTCCTAAGACTCAGATTGCAAAGGAAAAATTAGATCAG
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
D84223          GTTTTCTTCAAGGAGGCTCCATTTCCTAAGACTCAGATTGCAAAGGAAAAATTAGATCAG
                    1990      2000      2010      2020      2030      2040

                        520       530       540       550       560       570
33583           TTAAAGCAGGAGTTTGAATTCTGGTATCCTGTTGATCTTCGCGTCTCTGGCAAGGATCTT
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
D84223          TTAAAGCAGGAGTTTGAATTCTGGTATCCTGTTGATCTTCGCGTCTCTGGCAAGGATCTT
                    2050      2060      2070      2080      2090      2100
```

```
                580        590        600        610        620        630
33583       GTTCCAAATCATCTTTCATATTACCTTTATAATCATGTGGCTATGTGGCCGGAACAAAGT
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
D84223      GTTCCAAATCATCTTTCATATTACCTTTATAATCATGTGGCTATGTGGCCGGAACAAAGT
               2110       2120       2130       2140       2150       2160

                640        650        660        670        680        690
33583       GACAAATGGCCTACAGCTGTGAGAGCAAATGGACATCTCCTCCTGAACTCTGAGAAGATG
            ||  ||||||||||||||||||||||||||||||||||||||||||||||||||||||||
D84223      GATAAATGGCCTACAGCTGTGAGAGCAAATGGACATCTCCTCCTGAACTCTGAGAAGATG
               2170       2180       2190       2200       2210       2220

                700        710        720        730        740        750
33583       TCAAAATCCACAGGCAACTTCCTCACTTTGACCCAAGCTATTGACAAATTTTCAGCAGAT
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
D84223      TCAAAATCCACAGGCAACTTCCTCACTTTGACCCAAGCTATTGACAAATTTTCAGCAGAT
               2230       2240       2250       2260       2270       2280

                760        770        780        790        800        810
33583       GGAATGCGTTTGGCTCTGGCTGATGCTGGTGACACTGTAGAAGATGCCAACTTTGTGGAA
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
D84223      GGAATGCGTTTGGCTCTGGCTGATGCTGGTGACACTGTAGAAGATGCCAACTTTGTGGAA
               2290       2300       2310       2320       2330       2340

                820        830        840        850        860        870
33583       GCCATGGCAGATGCAGGTATTCTCCGTCTGTACACCTGGGTAGAGTGGGTGAAAGAAATG
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
D84223      GCCATGGCAGATGCAGGTATTCTCCGTCTGTACACCTGGGTAGAGTGGGTGAAAGAAATG
               2350       2360       2370       2380       2390       2400
```

I.M.A.G.E. cDNA clone 33583 has a 99.7% identity with the *Homo sapiens* mRNA for leucyl tRNA synthetase, over a 2263bp nucleotide overlap.

5.3.7 Summary

During this study, two new patients (patient 3 and patient 4) with MDS and a 5q deletion were identified which narrowed the critical region of the 5q- syndrome to approximately 3Mb at 5q31.3-q33 (patient 3), and then approximately 1.5Mb at 5q31.3-q32. The new critical region now excluded the *CDC60*, *RMSA-1*, and *PP2A* genes. No further analysis was carried out.

5.4 Discussion

The Human Chromosome 5 GeneMap and the UniGene set at NCBI identified six transcripts that mapped to the approximate 5Mb critical region of the 5q- syndrome at 5q31-q33 flanked by the genes *FGF1* and *IL12ß*. Each transcript was represented by ESTs 'highly similar' to known genes from yeast or *Drosophila*, or known human genes. These genes represented candidates for the 5q- syndrome tumour suppressor gene.

Three of the six genes (Goliath protein, *trp1*, and *PPP1R2*) were shown to map outside the critical region of gene loss by gene dosage analysis. These results, as with the novel genes in Chapter 4, highlight the redundancy in the EST database. The Goliath protein had been mapped by two independent groups at two locations on the Whitehead map, and five locations on the Transcript map between two sets of DNA markers. These multiple mappings indicated a discrepancy in its localisation. An updated UniGene search shows it has subsequently been localised to two regions at 5q35, outside the distal breakpoint of the 5q- syndrome at 5q33.

The *tpr1* gene had been mapped independently by Oxford using the GB4 panel. It is beneficial for a gene to be mapped to the same location by two or more independent groups to confirm the localisation. Murthy *et al.*, (1996) had previously localised the *tpr1* gene to chromosome 5q32-q33.2 flanked by the DNA markers D5S2049 and D5S1955. The novel human gene was identified by a two-hybrid screen when interacting with the GAP-related domain of neurofibromin, the product of the *NF1* gene. However, *tpr1* was found to map outside the distal breakpoint of the critical region, at 5q33, by gene dosage analysis. We subsequently mapped it to 5q33-q34 by gene dosage using the granulocyte DNA

from the newly identified patient 3. Recently, the Ensembl program has mapped the gene within contig AC000609 at 5q33.3.

The *PPP1R2* gene, like *trp1*, had been independently mapped using the GB4 panel. The gene was shown to map proximal to *trp1* between the DNA markers D5S410 and D5S487, according to the EST database. *PPP1R2* was also shown to map outside the distal breakpoint, but within 5q33-q34, by gene dosage analysis. The Ensembl program has mapped the gene within contig AC011414, also at 5q33.3, proximal to the *tpr1* gene. Therefore, the EST database had wrongly localised the two genes, but their position relative to each other was correct. Following these results, the *Homo sapiens* genome view showed the *PPP1R2* gene to have four "hits" in the human genome. Two "hits" were shown to map to 5q, while two were shown to map to 3q29, a region encoding many genes including the candidate tumour suppressor gene, *DLG1*, a human homologue of the *Drosophila* disc large tumour suppressor gene (Azim *et al.*, 1995). Subsequently, Permana and Mott (1997) determined the authentic *PPP1R2* gene to be located on chromosome 3q29 consisting of six exons, when investigating whether genetic alterations in *PPP1R2* could contribute to insulin resistance in Pima Indians. Permana and Mott showed the gene on chromosome 5 to be a homologue of *PPP1R2*, and identified it as an intronless pseudogene.

The *PP2A* gene was correctly mapped to the critical region of the 5q- syndrome at 5q31-q33, according to the GeneMap and UniGene set at NCBI. The *PP2A* gene had been mapped by two independent groups to the critical region of gene loss. Subsequently, the gene was shown to be retained in patient 3, therefore mapping the gene outside the new proximal breakpoint at 5q31.3.

Two transcripts were identified as the human homologues of the yeast cell division cycle gene *CDC60*, and the *Drosophila* regulator of mitotic spindle assembly, *RMSA-1* gene. Our data confirmed that these genes were correctly mapped to the critical region of the 5q- syndrome at 5q31-q33, according to the GeneMap and UniGene set at NCBI. Like the *PP2A* gene, the *CDC60* gene had been mapped by two independent groups to the critical region of gene loss, and then subsequently shown to be retained in patient 3, therefore mapping the gene outside the new proximal breakpoint at 5q31.3. However, the *RMSA-1* gene was shown to map to the new critical region of the 5q- syndrome at 5q31.1-q33 by gene dosage analysis. A subsequent search on the UniGene and the *Homo sapiens* genome view showed the gene to be localised to three regions on chromosome 5q, and one region on chromosome 11cen-q22.3. These multiple mappings indicated a discrepancy in the localisation of *RMSA-1*, therefore no further analysis was carried out. Since the end of this study, a new patient (patient 4) with MDS and a 5q deletion was identified, narrowing the critical region to approximately 1.5Mb at 5q31.3-q32. The Ensembl program has recently mapped *RMSA-1* within contig AC021078 at 5q33, thus outside the reduced critical region of the 5q- syndrome.

5.5 Conclusion

The EST resource was successfully used to localise one known human gene and identify two human homologues of known genes from *S.cerevisiae* and *D.melanogaster*, which mapped to the 5Mb critical region of gene loss in the 5q- syndrome, at 5q31-q33. These three genes thus represented potential candidate genes for the 5q- syndrome. However, during the course of this study, new patient data narrowed the 5q- syndrome critical region to approximately 1.5Mb at 5q31.3-q32, flanked by the genetic marker D5S413 and *GLRA1* gene. The three candidate genes have been excluded from the reduced critical region.

The main disadvantage of the EST resource is its high degree of redundancy. Fifty per cent of the six genes originally selected for analysis were shown to map outside the critical region of gene loss. This was previously highlighted in Chapter 4 when mapping novel genes to the 5q- syndrome critical region. The high failure rate may be due to redundancy in the sequence data, widely dispersed sequences, ambiguous nucleotides within the sequences, the possibility of amplifying through introns and the presence of repetitive elements within the sequence (Malone *et al.*, 1999). Therefore, gene dosage analysis was used to localise the genes to 5q and produce an accurate transcription map of the critical region of gene loss in the 5q- syndrome.

The process of generating a transcript map of the 5q- syndrome critical region has been facilitated by the increasing flow of data released by the Human Genome Project. In April 2000, researchers at the Department of Energy's Joint Genome Institute in California decoded in draft form the genetic information on human chromosome 5. Chromosome 5 contains an estimated 194 million bases, or about six percent of the human genome. Disease-linked genes on this chromosome

include those for colorectal cancer, basal cell carcinoma, acute myelogenous leukaemia, and a type of dwarfism. In addition, the approximate 1.5Mb critical region of the 5q- syndrome is currently being annotated by members of our group using the Ensembl program in collaboration with the Sanger Centre. The Ensembl program has used a number of criteria to map a particular gene to a particular chromosomal region, thereby decreasing the amount of redundancy in the database. For example, genes have only been predicted represented by ESTs that are part of a cluster in the UniGene set. ESTs represented as unidentified transcripts are not considered to be true genes. It is estimated that there are a total number of thirty-six genes, twenty-three known, and thirteen novel mapping to the critical region of the 5q- syndrome.

We have generated a detailed transcript map of the 5q- syndrome critical region comprising of known genes and novel coding sequences. Following the success of the HGP along with the annotation of the critical region of the 5q- syndrome, we will carry out mutation studies with the aim of identifying the 5q- tumour suppressor gene(s).

Chapter 6

Molecular analysis of the *SPARC*, *HAH1*, and *Annexin* VI genes

6.1 Introduction

6.1.1 Targeting genes as candidates for disease

6.1.2 Targeting genes in MDS and leukaemia

6.1.3 The *SPARC* gene

6.1.4 The *Annexin* VI gene

6.1.5 The role of antioxidants in cancer

6.1.6 The *HAH1* gene

6.2 Materials and Methods

6.2.1 Patients

6.2.2 Samples

6.2.3 Gene dosage analysis

6.2.4 Northern analysis

6.2.5 Southern analysis

6.2.6 Localisation to the YAC contig

6.2.7 Expression in $CD34^+$ cells by RT-PCR analysis

6.2.8 Mutation analysis of the *SPARC* gene

6.2.8.1 Polymerase Chain Reaction (PCR) amplification of the *SPARC* gene

6.2.8.2 Subcloning of *SPARC* exon PCR products

6.2.8.3 Sequencing of *SPARC* exon PCR products

6.2.9 Mutation analysis of the *HAH1* and *Annexin* VI genes

6.2.9.1 Reverse transcriptase PCR (RT-PCR)

6.2.9.2 Purification and quantification of RT-PCR products

6.2.9.3 Cycle sequencing on the ALF*express* automated sequencer

6.2.10 Database analysis using the Genetics Computer Group (GCG) software package

6.3 Results

6.3.1 Patients
6.3.2 Gene dosage analysis
6.3.3 Northern analysis
6.3.4 Southern analysis
6.3.5 Localisation to the YAC contig
6.3.6 Expression in CD34$^+$ cells by RT-PCR analysis
6.3.7 Mutation analysis of the *SPARC* gene
6.3.7 Mutation analysis of the *HAH1* gene
6.3.9 Mutation analysis of the *Annexin* VI gene
6.3.10 Database analysis using the Genetics Computer Group (GCG) software package

6.4 Discussion

6.4.1 The role of antioxidants in MDS and leukaemia
6.4.2 Future work

6.1 Introduction

6.1.1 Targeting genes as candidates for disease

Frequent deletions and loss of heterozygosity in a segment of a particular chromosome in association with a malignancy suggest that the disease may be caused by inactivation of a tumour suppressor gene located in the commonly deleted region (CDR). The identification and targeting of novel genes in such regions has followed many approaches including: constructing a high-resolution physical map of YAC, PAC, and cosmid contigs covering the genomic region, exon trapping, and direct selection as described in the previous chapter. With many CDRs being gene-rich, identifying candidate genes for mutation analysis requires the need for prioritisation.

Genes have been isolated as candidates due to their localisation, expression patterns, homology with a protein from another species, belonging to a gene family, and their predicted/known function. Many genes involved in carcinogenesis have been targeted due to their chromosomal localisation. The distal portion of chromosome 1p is one of the most commonly affected regions in human cancer (Chadwick *et al.*, 2000). A study of hereditary and sporadic colorectal cancer identified a region of frequent deletion 32.2 centimorgans (CM) from 1ptel (Chadwick *et al.*, 2000). Results showed deletion breakpoints to cluster in the vicinity of or inside the *RIZ* gene that encodes a retinoblastoma protein-interacting zinc finger protein. Moreover, the RIZ1 isoform contains a PR domain implicated in tumour suppressor function. The PR domain is a newly recognised protein motif that characterises a subfamily of Kruppel-like zinc finger genes. Members of the PR domain family have been shown to play important roles in cell differentiation and malignant transformation (Liu *et al.*, 1997). Other candidate tumour suppressor genes isolated as a result of their chromosomal localisation

include the *HIC-1* (Hypermethylated in cancer 1) a BTB/POZ transcriptional repressor located at 17p13.3, a region hypermethylated or subject to allelic loss in many human cancers (Guerardel *et al.*, 2001).

The technique of screening the human EST database has been used to target genes as candidates for disease. db(EST) has recently been used to identify candidate genes in respiratory chain deficiency. Disorders of mitochondrial oxidative phosphorylation (OXPHOS) are now recognised as major causes of human metabolic diseases and several mutations of mitochondrial and nuclear genes encoding respiratory chain components have been reported (Rotig *et al.*, 2000). While several hundred of these genes have been reported in yeast, only a few nuclear genes have been identified in humans. The yeast databases therefore present an invaluable tool for identification of human homologues that should be regarded as candidate genes in OXPHOS diseases. Yeast protein sequences were compared to the GenBank db(EST) database in order to identify the human counterparts, using the BLAST program. The study identified one hundred and two groups of human ESTs with significant homology to yeast genes (Rotig *et al.*, 2000).

6.1.2 Targeting genes in MDS and leukaemia

The elucidation of the putative tumour suppressor gene(s) involved in the pathogenesis of many myeloid and lymphoid malignancies remains a major goal in cancer research. Several genes with tumour suppressor activity have been targeted as candidates for MDS and leukaemia. *p53* is one of the most frequently mutated genes in human cancers (Neubauer *et al.*, 1993). Since *p53* has been implicated in lymphatic and some myeloid leukaemias, such as the blastic phase of CML, studies have been carried out to address the role of *p53* gene mutations in

MDS. The study by Neubauer *et al.*, looked at mutations within exons 4-9 of the *p53* gene, in patients with MDS. No mutations were found in the seventeen MDS patients included in the study suggesting *p53* gene mutations do not play a major role in the pathogenesis of MDS. However, a study by Kaneko *et al.*, (1995) looked at exons 5-8 of the *p53* gene. *p53* mutations were found in 7/57 (12%) patients with MDS. These mutations correlated with both leukaemic transformation and a poor prognosis in MDS.

We decided to target known genes as candidates for the 5q- syndrome. Due to the speed and accuracy of the EST database, we used the technique of screening db(EST) for the identification of candidate tumour suppressor genes. We have analysed the *SPARC* gene which has been reported to possess tumour suppressor activity (Mok *et al.*, 1996); the *annexin* VI gene which belongs to a family of genes of which have been implicated in tumourigenesis (Kataoka *et al.*, 2000); and the *ATOX1* (previously *HAH1*) gene which has been predicted to play a role in antioxidant defence (Klomp *et al.*, 1997).

6.1.3 The *SPARC* gene

SPARC (secreted protein acidic and rich in cysteine), also termed BM-40, 43K-protein, and osteonectin is an extracellular, Ca2+ ion-binding, glycoprotein widely distributed in human tissues undergoing developmental regulation and repair (Lane and Sage, 1994; Ledda *et al.*, 1997). The protein is shown to be expressed in haematopoietic cells including megakaryocytes (Kelm Jr *et al.*, 1992). Although information concerning the expression, biochemical properties, and cellular activities of *SPARC* has increased significantly over the last fifteen years, the precise function of the protein is unknown. However, a number of studies have shown *SPARC* to possess important properties. One such property is as an antiadhesin causing changes in cell shape by disrupting cell-matrix interactions

(Lane and Sage, 1994). *SPARC* has also been shown to function as a cell growth regulator by inhibiting progression through the $G_1 \rightarrow S$ phase of the cell cycle (Sage and Bornstein, 1991). These functions suggest *SPARC* may play a role in tumourigenesis. Moreover, Mok *et al.*, (1996) showed that downregulation of *SPARC* strongly reduced the growth rate of cancer cell lines *in vitro*, suggesting *SPARC* functions as a tumour suppressor gene, at least in some malignancies.

The *SPARC* gene has been localised to chromosome 5q31-q33 by somatic cell hybrid analysis and by *in situ* hybridisation (Swaroop *et al.*, 1988); within the critical region of gene loss in the 5q- syndrome (Boultwood *et al.*, 1994; Jaju *et al.*, 1998), see Figure 6.1. Its expression in haematopoietic cells along with its properties as a cell-cycle inhibitor make *SPARC* a candidate for the putative tumour suppressor gene associated with the 5q- syndrome. The *SPARC* gene was analysed for mutations by direct sequencing in three patients with MDS and the 5q- syndrome.

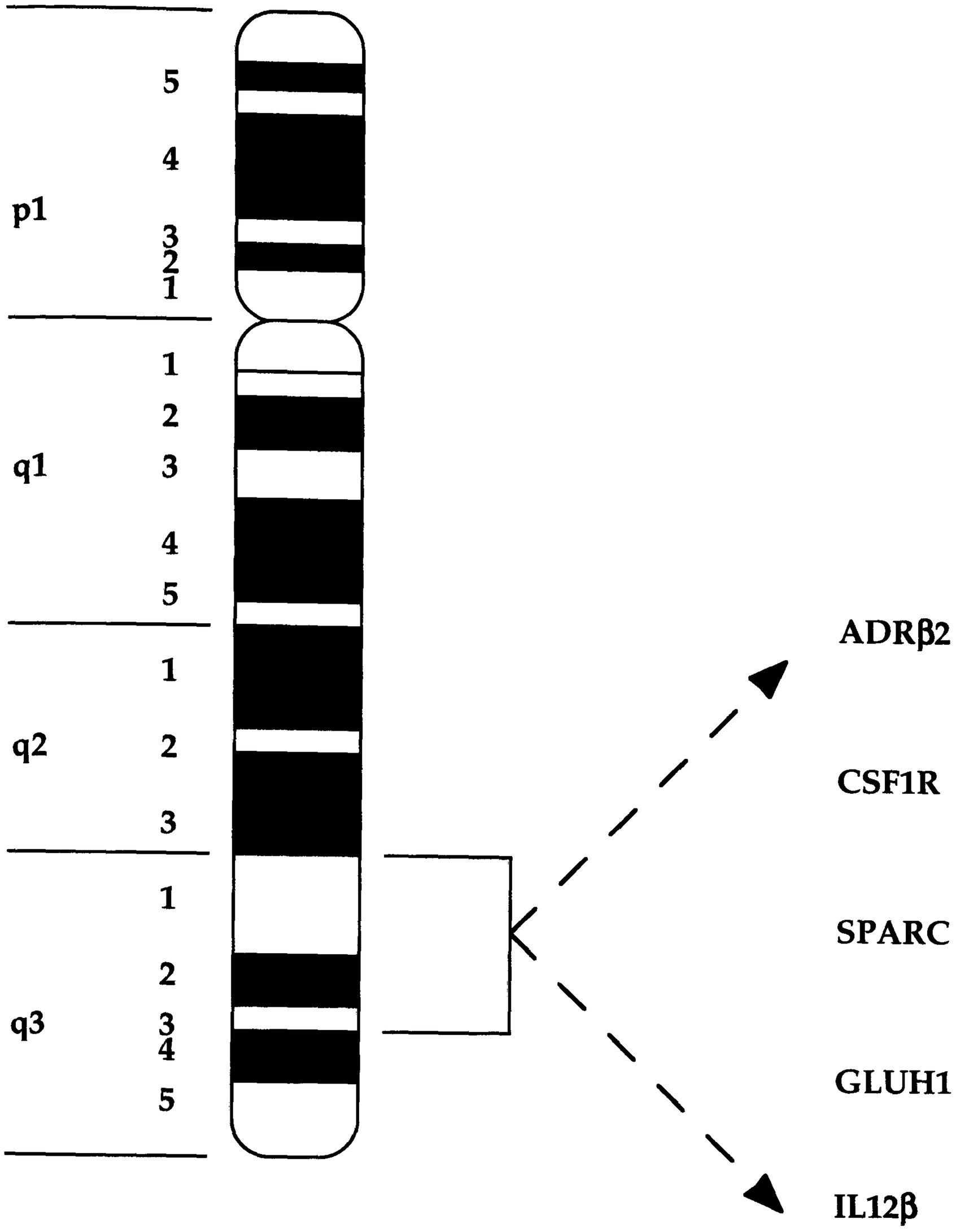


Figure 6.1 Ideogram of Chromosome 5 illustrating the critical region of the 5q- syndrome and the position of the *SPARC* gene

6.1.4 The *Annexin* VI gene

The annexins were first described as a family of calcium phospholipid-binding proteins in 1990 (Crumpton and Dedman, 1990). To date, over twenty members of the annexin family have been identified, ten of which have been described in mammals, including *annexin* VI. The annexins are defined by a conserved internally repetitive 70-amino acid sequence, present four times in all annexins except *annexin* VI, which has eight repeats (Crompton *et al.*, 1988). Although the

biological functions of these proteins have yet to be established, *annexin* V has been proposed to play a role in apoptosis (Rand, 1999). The fact that they are highly conserved and present in a variety of cell types suggests that important biological roles will be elucidated. *Annexin* VI is highly expressed in mammalian tissues although generally restricted to specialised cell types, in particular endocrine cells and certain ductal epithelia (Clark *et al.*, 1991). This pattern of expression supports a role for *annexin* VI in some aspect of secretion, although *in vitro* studies showed *annexin* VI to inhibit *synexin* (*annexin* VII) and *calpactin* (*annexin* II) driven granule aggregation (Creutz *et al.*, 1992).

Although no one clear function for the annexins has been established, a number of studies suggest a role in tumour suppression. A study of the *annexin* VI expression levels on a melanoma cell line showed the gene to be a marker for the less invasive phenotype of malignant melanoma, and suggested a possible role in tumour suppression (Kataoka *et al.*, 2000). Moreover, *annexin* VI, like the *SPARC* gene, has been shown to inhibit progression through the cell cycle, suggesting a possible role for *annexin* VI in cell growth regulation and tumourigenesis (Theobald *et al.*, 1994). A subsequent study showed the heterologous expression of *annexin* VI in A431 squamous carcinoma cells caused a marked suppression of tumour cell growth (Theobald *et al.*, 1995).

The *annexin* VI gene has been localised to the critical region of gene loss in the 5q- syndrome by gene dosage analysis, and sublocalised to the YAC contig (Boultwood *et al.*, 2000). This localisation, expression pattern, and properties of *annexin* VI as a cell cycle inhibitor make it a candidate for the putative tumour suppressor gene associated with the development of the 5q- syndrome. The *annexin* VI gene was analysed for mutations by cycle sequencing in nine patients with MDS and the 5q- syndrome.

6.1.5 The role of antioxidants in cancer

Genes having known or predicted tumour suppressor activity, for example the *SPARC* and *annexin* VI genes respectively, represent good candidate genes for the 5q- syndrome. Also, genes with a predicted function in antioxidant defence have been shown to be involved in the pathogenesis of cancer. A gene localised to the 5q- syndrome critical region predicted to be involved in antioxidant defence is the *ATOX1* (previously *HAH1*) gene.

Reactive oxygen species are widely generated in biological systems. Consequently, humans have evolved antioxidant defence systems that limit their production. Intracellular production of active oxygen species such as dioxygen and hydrogen peroxide is associated with the arrest of cell proliferation. Similarly, generation of oxidative stress in response to various external stimuli has been implicated in the activation of transcription factors and to the triggering of apoptosis. Despite antioxidant defence mechanisms, cell damage from oxygen free radicals (OFRs) is ubiquitous. OFR-related lesions that do not cause cell death can stimulate the development of cancer (Dreher and Junod, 1996). Reducing the avoidable endogenous and exogenous causes of oxidative stress is the current strategy at present, but in the future, the action of tumour suppressor genes and the DNA repair mechanisms may lead the way to additional tools against carcinogenesis from OFR (Dreher and Junod, 1996).

6.1.6 The *HAH1* gene

The human ATX homologue 1 (*HAH1*) gene is the human homologue of the *ATX1* gene in *Saccharomyces cerevisiae* (Klomp *et al.*, 1997). *ATX1* encodes a cystolic copper-binding protein in *S. cerevisiae*, functioning to protect cells from toxicity in a copper-dependent manner (Klomp *et al.*, 1997). The *ATX1* protein (Atx1p) functions as an antioxidant protecting yeast from the toxic effects of superoxide

and hydrogen peroxide (Lin and Culotta, 1995). Therefore, it was suggested that *HAH1* may play an essential role in the antioxidant defence and copper homeostasis in humans (Klomp *et al.*, 1997).

Fluorescence *in situ* hybridisation had previously localised *HAH1* to chromosome 5q32-q33 (Klomp *et al.*, 1997). It was recently finely mapped to the YAC contig encompassing the critical region of the 5q- syndrome, adjacent to the *SPARC* gene (Boultwood *et al.*, 2000). It's localisation, expression pattern, and properties as an antioxidant make *HAH1* a candidate for the tumour suppressor gene associated with the 5q- syndrome. The *HAH1* gene was analysed for mutations by cycle sequencing in eight patients with MDS and the 5q- syndrome.

6.2 Materials and Methods

6.2.1 Patients

Three patients with MDS and a 5q deletion were included in the direct sequencing study of the *SPARC* gene. A further 6 patients were included in the molecular analysis and cycle sequencing study of the *HAH1* and *annexin* VI genes. Classification was according to the FAB criteria (Kouides and Bennett, 1992). At the time of investigation, all 9 patients had the characteristic clinical and haematological features of the 5q- syndrome (Van den Berghe *et al.*, 1974; Dewald *et al.*, 1985)

6.2.2 Samples

Mononuclear cells and granulocytes were separated from 40mls of EDTA treated peripheral blood by Ficoll gradient centrifugation (Boyum, 1984). The granulocyte fraction showed a high level of purity (≥95%). High molecular weight DNA was obtained from the fractionated peripheral blood by Nucleon® extraction. Granulocyte DNA from one healthy individual was used as the control.

Total RNA was obtained from the patient granulocyte fractions with the Totally RNA Isolation Kit that is based on the disruption of cells in guanidinium thiocynate/cationic detergent solutions, followed by organic extraction and alcohol precipitation of the RNA. Granulocyte total RNA fractions from the peripheral blood of healthy individuals were used as controls.

6.2.3 Gene dosage analysis

The *SPARC* gene had previously been localised to the critical region of the 5q- syndrome by gene dosage analysis (Boultwood *et al.*, 1994). cDNA clones representing; (1) part of the *HAH1* gene, and (2) the *annexin* VI gene were localised

to the critical region of gene loss in this study as previously described (**Chapter 3 section 3.2.4**).

6.2.4 Northern analysis

An I.M.A.G.E. cDNA clone representing part of the *HAH1* gene was hybridised to an MTN blot as previously described (**Chapter 3 section 3.2.5**), to determine the tissue expression pattern and transcript size of the cDNA. Northern analysis had previously been carried out on the *SPARC* gene (Xavier *et al.*, 1989) and the *annexin* VI gene (Smith *et al.*, 1994).

6.2.5 Southern analysis

Granulocyte DNA fractions were obtained from nine 5q- syndrome patients and healthy individuals. The DNA was digested with restriction enzymes *Pst*I, *Hind*III, *Pvu*II, *and Bgl*II; size fractionated through a 1% agarose gel and Southern blotted. Two Southern blot filters were prepared and hybridised with an I.M.A.G.E. cDNA clone insert derived from the *HAH1* gene to screen for gene rearrangements. Southern analysis was carried out on two separate occasions. Southern analysis had previously been carried out on the *SPARC* and *annexin* VI genes in the laboratory.

6.2.6 Localisation to the YAC contig

The *SPARC, HAH1,* and *annexin* VI genes were sublocalised by PCR screening to the YAC contig encompassing the critical region of the 5q- syndrome (Kostrzewa *et al.*, 1998) as previously described (**Chapter 2 section 2.15**). The genes were further localised to BACs (Bacterial artificial chromosomes) spanning the critical region of gene loss. PCR primer pairs were designed from the coding region of each gene. Details of the primer conditions are shown in Table 6.1.

Table 6.1 **PCR primer conditions for localisation of the *SPARC*, *HAH1*, and *annexin* VI genes to the YAC contig**

Gene	Primer name	Primer sequence 5'-3'	Annealing temp	PCR size
SPARC	SPARC-F9	TTCCCTGCAGGTACCTCT	55°C	150bp
	SPARC-R9	ACTCACTCTGCTTGATGC		
HAH1	HAH1-FB	GAGGCGCTGCTGACAC	61°C	416bp
	HAH1-RC	CAACAAAAGCAGCTTGATTTATG		
Annexin VI	ANX6-F1	GCACTTCTGCCAAGAAATGG	58°C	320bp
	ANX6-R1	ACAGACAGAGGTTCAGGATG		

6.2.7 Expression in CD34$^+$ cells by RT-PCR analysis

CD34$^+$ expression analysis was carried out on the *SPARC*, *HAH1*, and *annexin* VI genes as previously described (Chapter 2 section 2.17). RT-PCR primer pairs were designed flanking the coding regions of each gene. Details of the primer conditions are shown in Table 6.2.

Table 6.2 **PCR primer conditions for CD34$^+$ expression analysis in the *SPARC*, *HAH1*, and *annexin* VI genes**

Gene	Primer name	Primer sequence 5'-3'	Annealing temp	PCR size
SPARC	SPARC-RTF1	AAATACATCCCCCCTTGCC	56°C	557bp
	SPARC-RTR1	CAGAACAACAAACCATCCAAAC		
HAH1	HAH1-FB	GAGGCGCTGCTGACAC	61°C	440bp
	HAH1-RC	CAACAAAAGCAGCTTGATTTATG		
Annexin VI	ANX6-F4	GATGCTCTGAGCTCAGACAC	60°C	700bp
	ANX6-R4	AGATAAGAGCCCAACCCAAC		

6.2.8 Mutation analysis of the *SPARC* gene

The method of choice for mutation analysis of the *SPARC* gene was directsequencing of the 9 coding exons since the genomic structure of the gene was known.

6.2.8.1 Polymerase Chain Reaction (PCR) amplification of the *SPARC* gene

PCR amplification was performed on patient granulocyte DNA samples and on the normal granulocyte DNA from one healthy individual. Primer pairs were designed from intronic sequences flanking each of the 9 coding exons (exons 2-10) of the *SPARC* gene. Details of the primer conditions are shown in Table 6.3. PCR was performed on a thermal cycler (Biometra Trio Thermoblock) in a 50µl reaction volume in PCR buffer containing 1mM-3mM Mg^{2+} (Table 6.3), 1.25µM dNTPs, 2.5U of *Taq* polymerase, 200ng of template DNA, and 100pmoles of each of the primers for 35 cycles under the following conditions: 94°C for 30 seconds, annealing temperature (Table 6.3) for 30 seconds, and 72°C for 1 minute. The PCR products were run on 1-2% agarose gels (dependent on size of PCR product) and purified using Wizard® PCR Preps DNA Purification System (Promega) as previously described (**Chapter 2 section 2.5.1.3)**

Table 6.3 ***SPARC*** **exon PCR conditions**

SPARC coding exon	Primer name	Primer sequence 5'-3'	Mg^{2+}	Annealing temperature	Product size
2	SPARC-2F SPARC-2R	GTTCCCAGCACCATGAGGGC ACTTACAGGGGCTGCCAA	1mM	60°C	68bp
3	SPARC-3F SPARC-3R	CACTAGCAGCAAGAAGCC ACATACCTCAGTCACCTC	1mM	58°C	146bp
4	SPARC-4F SPARC-4R	TTCCAGGTATCTGTGGGA ACATACTTTCCGCCACCA	3mM	62°C	99bp
5	SPARC-5F SPARC-5R	CTACAGATCCCTGCCAGA CCTCACCTTCTCAAACTC	3mM	62°C	121bp
6	SPARC-6F SPARC-6R	CAACAGGTGTGCAGCAAT ACTCACATTTGCAAGGCC	3mM	63°C	122bp
7	SPARC-7F SPARC-7R	ACCTAGACATCCCCCCTT ACTTACCCGCAGCTTCTG	3mM	62°C	143bp
8	SPARC-8F SPARC-8R	CCTCAGGTGAAGAAGATC TCTTACCCGTCAATGGGG	3mM	55°C	149bp
9	SPARC-9F SPARC-9R	TTCCCTGCAGGTACCTCT ACTCACTCTGCTTGATGC	3mM	55°C	149bp
10	SPARC-10F SPARC-10R	TTGCAGAGGATATCGACA ACATTGTTAGCACCTTGT	3mM	55°C	142bp

6.2.8.2 Subcloning of *SPARC* exon PCR products

The purified PCR products were cloned into the pGEM®-T Easy Vector System II (Promega) as previously described (**Chapter 2 section 2.5.1.4**). The plasmid DNA was extracted using the QIAprep® Spin Miniprep Kit as previously described (**Chapter 2 section 2.5.1.2.**).

6.2.8.3 Sequencing of *SPARC* exon PCR products

Plasmid DNA was sequenced as a double-stranded template using the Cy5 Autoread sequencing kit as previously described (**Chapter 2 section 2.12.1.3-2.12.1.4**).

A minimum of 10 clones from each exon were sequenced from each patient and control.

6.2.9 Mutation analysis of the *HAH1* and *Annexin* VI genes

In contrast to the *SPARC* gene, only the cDNA sequence of the *HAH1* and *annexin* VI genes were known. The method of choice for mutation analysis of the coding regions of the *HAH1* and *annexin* VI genes was RT-PCR followed by cycle sequencing. *HAH1* has a small coding region (207bp) which could be cycle sequenced in one fragment fragment only. The *annexin* VI gene has a large coding region of 2022bp but was split into 4 fragments (approximately 500bp in size) and each fragment sequenced individually.

6.2.9.1 Reverse transcriptase PCR (RT-PCR)

RT-PCR was carried out using the R*everse*-iT™ One-step PCR kit as previously described (**Chapter 2 section 2.16**). Primer pairs were designed flanking the

coding regions of *HAH1* and *annexin* VI. Details of the primer conditions are shown in Table 6.4.

6.2.9.2 Purification and quantification of RT-PCR products

Purification and quantification of each patient and control RT-PCR product from each gene was carried out as previously described (**Chapter 2 section 2.16.1**).

6.2.9.3 Cycle sequencing on the ALF*express* automated sequencer

All cycle sequencing reactions were carried out using the Thermo SequenaseCy™5 Dye Terminator Kit (Amersham Pharmacia Biotech, Uppsala, Sweden), for use on the Alf*express* automated sequencer as previously described (**Chapter 2 section 2.18**). Nested primers were used for sequencing of the coding regions of the *HAH1* and *annexin* VI genes. Nested primers add specificity to the annealing and sequencing reactions. The increase in specificity results from the nested primer not annealing to any primer dimers or oligomers created in the PCR reaction. The primers were diluted in distilled water to a final concentration of 2pmol/µl. Details of the primers are shown in Table 6.5.

6.2.10 Database analysis using the Genetics Computer Group (GCG) software package

Sequence data generated from each patient and control of the three candidate genes: *SPARC, HAH1,* and *annexin* VI, was compared with the published sequences using BestFit analysis, as previously described (**Chapter 3 section 3.2.14.1**).

Table 6.4 ***HAH1*** **and** ***annexin*** **VI RT-PCR conditions**

Gene	cDNA fragment	Primer name	Primer sequence 5'-3'	Annealing temperature	RT-PCR product size
HAH1	1	HAH1-FB	GAGGCGCTGCTGACAC	61°C	416bp
		HAH1-RC	CAACAAAAGCAGCTTGATTTATTG		
Annexin VI	1	ANX6-F1	TTGCTGCTGGGCTAACGG	60°C	713bp
		ANX6-R1	GAACTGGGCTTCATCTGTTC		
	2	ANX6-F2	TCCAGAAGATGCTTGTGGTC	60°C	638bp
		ANX6-R2	GTCAGGGTTGAAGTCATTGG		
	3	ANX6-F3	AGAAGACTCTGCTGAAGCTG	60°C	629bp
		ANX6-R3	AAGATCTCAGCAGCCACCTG		
	4	ANX6-F4	GATGCTCTGAGCTCAGACAC	60°C	700bp
		ANX6-R4	AGATAAGAGCCCAACCCAAC		

Table 6.5 ***HAH1*** **and** ***annexin*** **VI cycle sequencing PCR conditions**

Gene	cDNA fragment	Primer name	Primer sequence 5'-3'	Annealing temperature
HAH1	1	HAH1-N1	GAGAGGCGCTGCTGACACC	63°C
		HAH1-N2	CACACCGCCGCTGCCTCAG	63°C
		HAH1-N3	GTCCATCCTGTGGGCTGTG	61°C
		HAH1-N4	CCAGGTCTGTCTGGAAGCC	61°C
Annexin VI	1	ANX6-N1	GCGGCTGGATTCTGCTGCG	62°C
		ANX6-N2	GCCTCGTATAGGTCCTGGAC	62°C
	2	ANX6-N3	GGAGGATGACGTAGTGAGCG	64°C
		ANX6-N4	AGCTCTACTCGGGCCACTG	61°C
		ANX6-N9	CTCCGTGAGGTGATTATGTCC	63°C
	3	ANX6-N5	GCAGCGCAGGTGGCCTATC	63°C
		ANX6-N6	CTCCTCACGATGCCCCGTG	61°C
		ANX6-N10	TTCCCGTGCCTGGTCCAGG	63°C
	4	ANX6-N7	CTGGCCACTTCAGGAGGATC	61°C
		ANX6-N8	TCAGGCTTGGCCATGGCGG	63°C

6.3 Results

6.3.1 Patients

Nine patients with the classical features of the 5q- syndrome and a deletion of chromosome 5 as the sole karyotypic abnormality were included in the molecular and mutation analysis of the *SPARC*, *HAH1*, and *annexin* VI candidate genes. Cytogenetic details of the patients are shown in Table 6.6.

6.3.2 Gene dosage analysis

Gene dosage analysis with cDNA clones derived from the *HAH1* and *annexin* VI genes and the *renin* gene showed both probes hybridised to a single fragment. An approximate 50% reduction in the dosage of each cDNA clone in the granulocyte patient DNA compared with normal controls, confirmed the deletion of one allele.

6.3.3 Northern analysis

Northern analysis showed I.M.A.G.E. cDNA clone 416547 derived from the *HAH1* gene to possess a single transcript of 0.8kb and to be ubiquitously expressed, see Figure 6.2.

6.3.4 Southern analysis

No rearrangements were observed in the granulocyte fractions of nine patients with the 5q- syndrome digested with restriction enzymes *Pst*I, *Hind*III, *Pvu*II, and *Bgl*II, following hybridisation with I.M.A.G.E. cDNA clone 1883868 derived from the *HAH1* gene.

Table 6.6 Clinical details of 5q- syndrome patients included in the study

Patient	Sex/age	Cytogenetic karyotype	Sample type
1	F/66	46, XX, del(5)(q31q33)	Granulocyte fraction DNA + RNA
2	F/22	46, XX, del(5)(q31q33)	Granulocyte fraction DNA + RNA
3	F/65	46, XX, del(5)(q33-q34)	Granulocyte fraction DNA + RNA
4	F/60	46, XX, del(5)(q13-q33)	Granulocyte fraction RNA
5	F/81	46, XX, del(5)	Granulocyte fraction RNA
6	M/48	46, XY, del(5)(q13-q33)	Granulocyte fraction RNA
7	F/78	46, XX, del(5)	Granulocyte fraction RNA
8	F/61	46, XX, del(5)(q13-q33)	Granulocyte fraction RNA
9	F/83	46, XX, del(5)(q13-q33)	Granulocyte fraction RNA

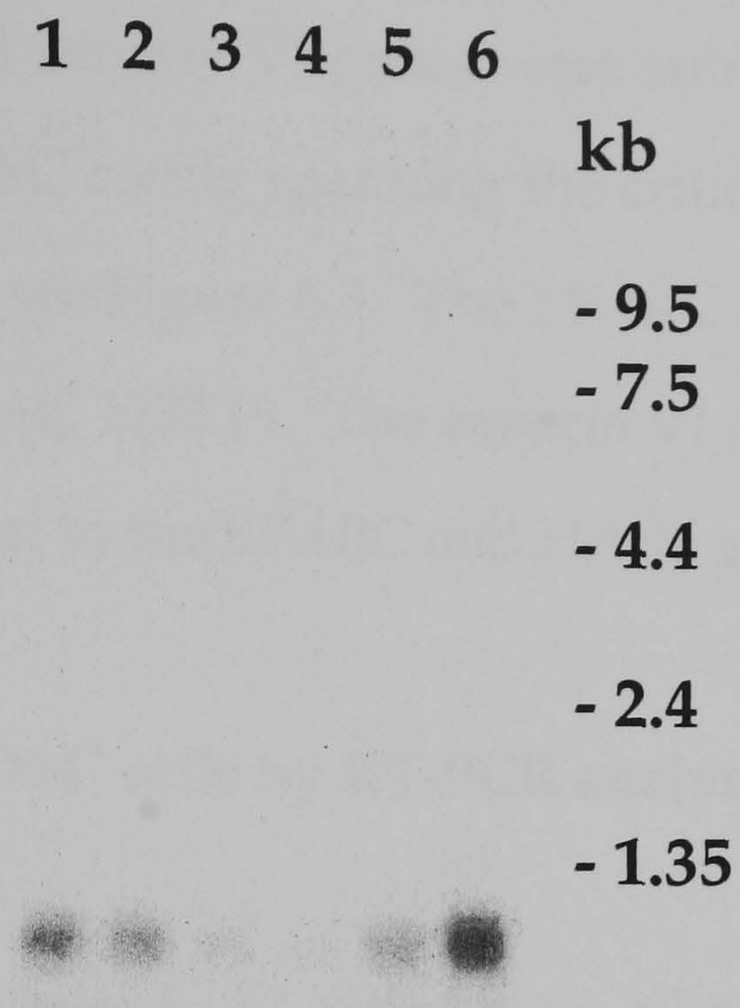


Figure 6.2

Representative Northern blot analysis of I.M.A.G.E. cDNA clone 416547. The MTN blot included 2µg of Poly-(A)+ RNA from; spleen (1), lymph node (2), thymus (3), peripheral blood leukocytes (4), bone marrow (5), and foetal liver (6). Sizes of RNA marker bands (kb) are indicated approximately.

6.3.5 Localisation to the YAC contig

The *SPARC*, *HAH1*, and *annexin* VI genes were sublocalised by PCR screening to YAC 816D6 from the YAC contig spanning the critical region of the 5q- syndrome (Kostrzewa *et al.*, 1998), see Figure 6.3. The *SPARC* and *HAH1* genes were further localised to the 200kb BAC 113P19. The *annexin* VI gene was shown to map to the 136kb BAC 119C17, distal to the *SPARC* and *HAH1* genes, at 5q32.

6.3.6 Expression in $CD34^+$ cells by RT-PCR analysis

The *SPARC*, *HAH1* and *annexin* VI genes were all expressed in RNA from $CD34^+$ cells upon analysis by RT-PCR, see Figure 6.4.

6.3.7 Mutation analysis of the *SPARC* gene

No mutations were found in the 9 coding exons (exons 2-10) of the *SPARC* gene in the 3 patients with the 5q- syndrome included in the study. A C to G substitution in exon 10 of patient 3 was observed at nucleotide 998. This single base substitution was also observed in the normal control, suggesting this was a polymorphism, see Figure 6.5. Patients 1 and 2 had a C at nucleotide 998, consistent with the published sequence (Villarreal *et al.*, 1989). Furthermore, in exon 10, a T was observed at nucleotide 960 in all 3 patients and the normal control. This is in contrast to the published sequence (Villarreal *et al.*, 1989) that has a G at nucleotide 960. This base change does not affect the amino acid sequence at this position.

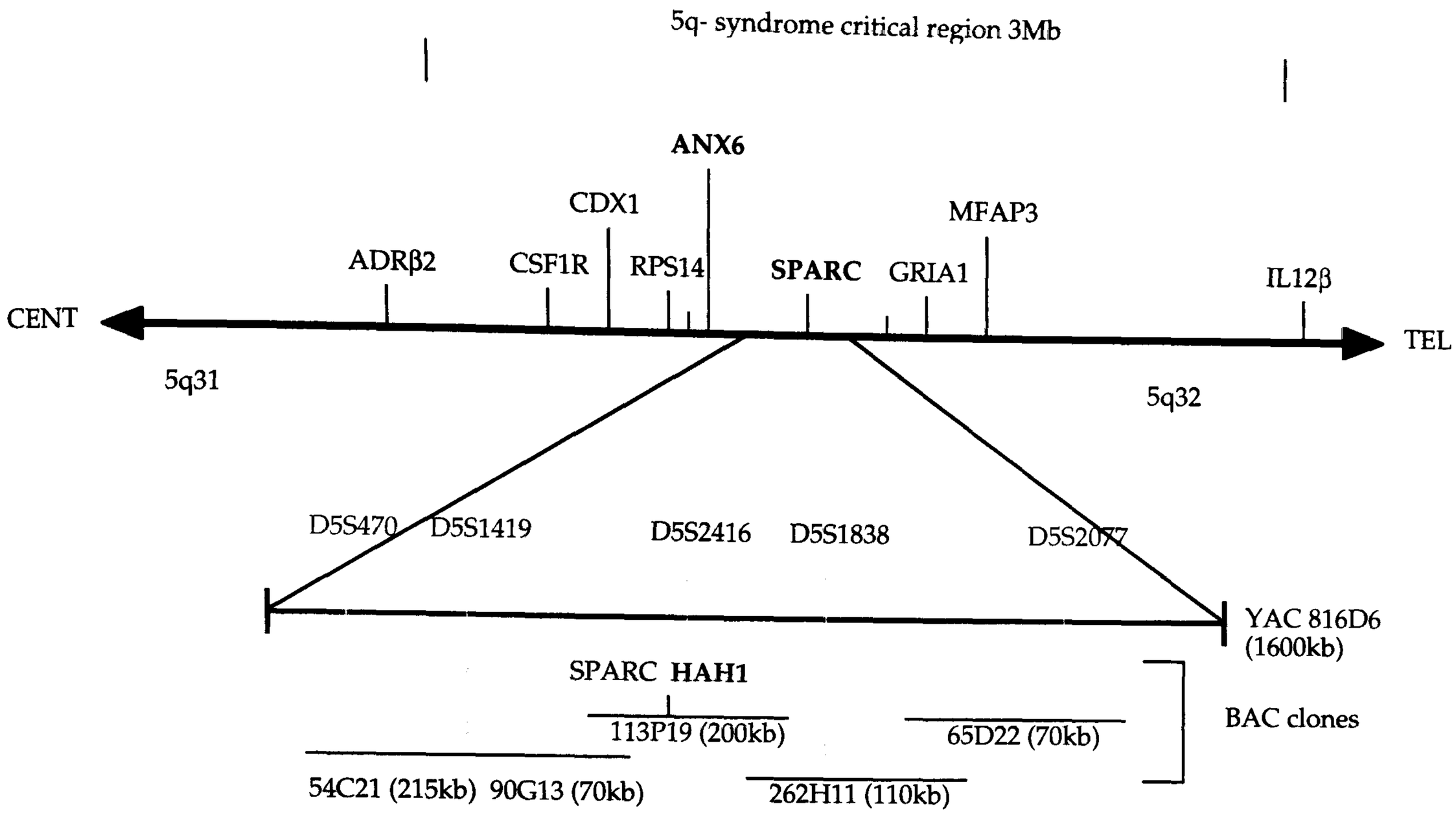


Figure 6.3

Fine physical mapping of the *SPARC, HAH1* and *annexin* VI (*ANX6*) genes (boldface) within the critical region of the 5q- syndrome at chromosome 5q32. All three genes were localised to YAC 816D6. The *SPARC* gene was sublocalised to BAC 113P19, adjacent to the *HAH1* gene. The *ANX6* gene was sublocalised to BAC 119C17. The size of the YAC and BACs shown is indicated in the brackets.

6.3.8 Mutation analysis of the *HAH1* gene

No mutations were observed in the 207bp coding region of the *HAH1* gene in the 8 patients with the 5q- syndrome included in the study.

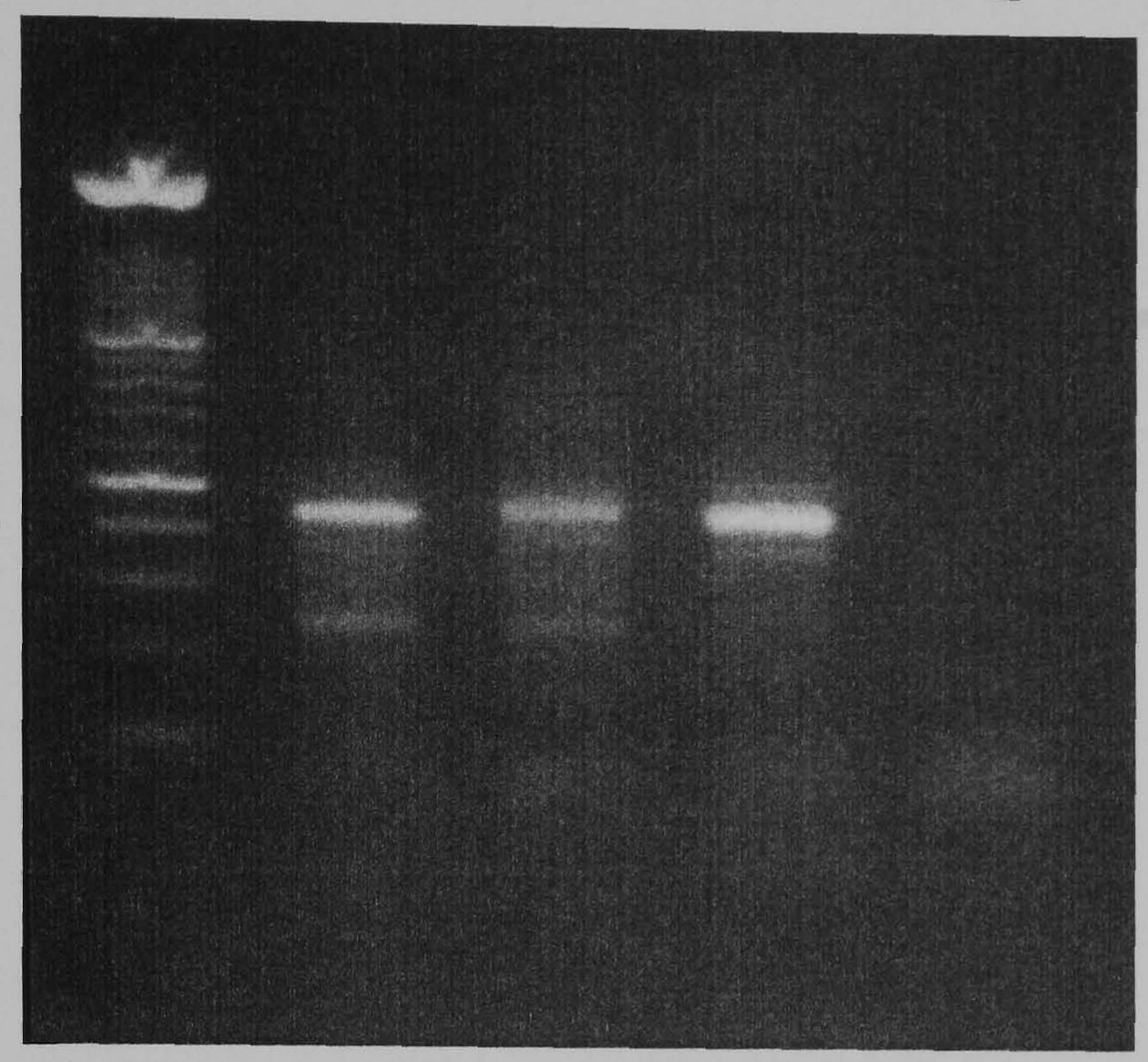
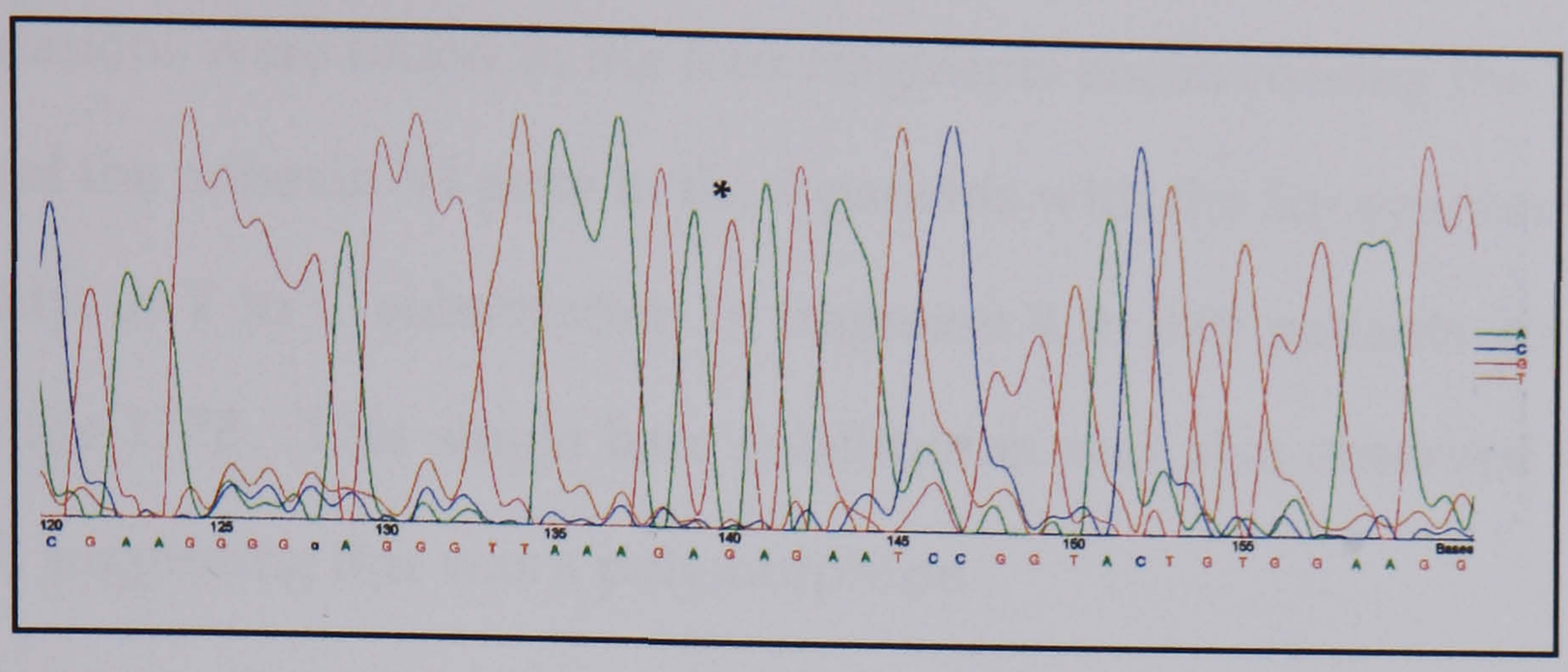


Figure 6.4

Representative $CD34^+$ expression analysis of the *annexin* VI gene. Total RNA obtained from $CD34^+$ cells 1:10 dilution (lane **1**), 1:100 dilution (lane **2**), and mononuclear cells of a normal healthy control (lane **3**) was amplified with *annexin* VI gene specific primers. A negative control (lane **4**) was included in the experiment.

a)



b)

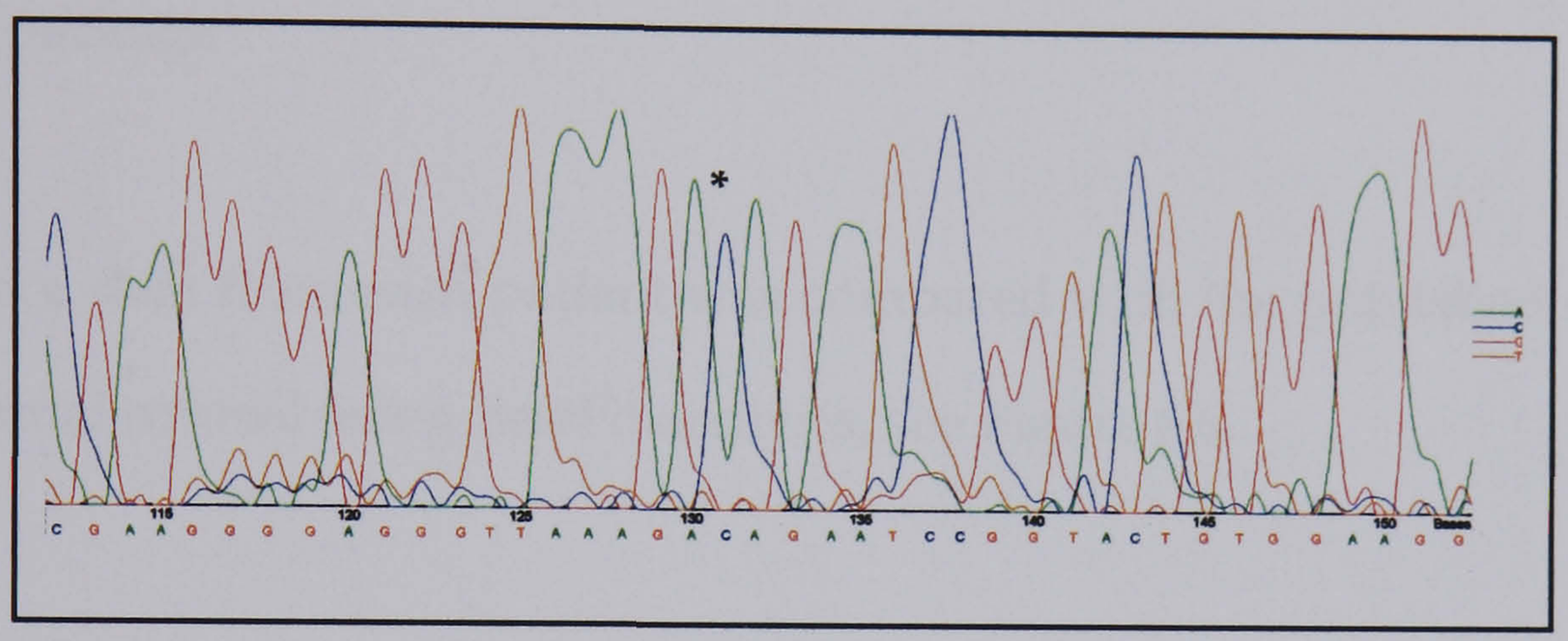


Figure 6.5

Representative sequence analysis of exon 10 of the *SPARC* gene. A C to G substitution in patient 3 (a) was observed at nucleotide 998. This single base substitution was also observed in the normal control suggesting this was a polymorphism. Patient 2 (b) shows a C at nucleotide 998, consistent with the published sequence. An asterisk (*) indicates the position of the base change.

6.3.9 Mutation analysis of the annexin VI gene

No mutations were found in the four fragments encompassing the 2022bp coding region of the annexin VI gene in the 9 patients with the 5q- syndrome included in the study. A T to C substitution in fragment 4 in 6/9 patients was observed at nucleotide 1778. This single base substitution was also observed in the normal control, suggesting this was a polymorphism.

6.3.10 Database analysis using the Genetics Computer Group (GCG) software package

Sequence data from each patient was compared with the published sequence and the normal control using BestFit analysis, see Figure 6.6.

```
   1 aggacctggtacaacaggatgtccaggacctatacgaggcaggggaactg 50
     ||||||||||||||||||||||||||||||||||||||||||||||||||
 691 aggacctggtacaacaggatgtccaggacctatacgaggcaggggaactg 740

  51 aaatggggaacagatgaagcccagttcatttacatcttgggaaatcgcag 100
     ||||||||||||||||||||||||||||||||||||||||||||||||||
 741 aaatggggaacagatgaagcccagttcatttacatcttgggaaatcgcag 790

 101 caagcagcatcttcggttggtgttcgatgagtatctgaagaccacaggga 150
     ||||||||||||||||||||||||||||||||||||||||||||||||||
 791 caagcagcatcttcggttggtgttcgatgagtatctgaagaccacaggga 840

 151 agccgatgaaggccagcatccgaggggagctgtctggggactttgagaag 200
     ||||||||||||||||||||||||||||||||||||||||||||||||||
 841 agccgatgaaggccagcatccgaggggagctgtctggggactttgagaag 890

 201 ctaatgctggccgtagtgaagtgtatccggagcaccccggaatattttgc 250
     ||||||||||||||||||||||||||||||||||||||||||||||||||
 891 ctaatgctggccgtagtgaagtgtatccggagcaccccggaatattttgc 940

 251 tgaaaggctcttcaaggctatgaagggcctggggactcgggacaacaccc 300
     ||||||||||||||||||||||||||||||||||||||||||||||||||
 941 tgaaaggctcttcaaggctatgaagggcctggggactcgggacaacaccc 990

 301 tgatccgcatcatggtctcccgtagtgagttggacatgctcgacattcgg 350
     ||||||||||||||||||||||||||||||||||||||||||||||||||
 991 tgatccgcatcatggtctcccgtagtgagttggacatgctcgacattcgg 1040

 351 gagatcttccggaccaagtatgagaagtccctctacagcatgatcaagaa 400
     ||||||||||||||||||||||||||||||||||||||||||||||||||
1041 gagatcttccggaccaagtatgagaagtccctctacagcatgatcaagaa 1090

 401 tgacacctctggcgagtacaagaagactctgctgaagctgtctgggggag 450
     ||||||||||||||||||||||||||||||||||||||||||||||||||
1091 tgacacctctggcgagtacaagaagactctgctgaagctgtctgggggag 1140

 451 atgatgatgctgctggccagtt 472
     |||||||||||||||||||||||
1141 atgatgatgctgctggccagtt 1162
```

Figure 6.6

Representative BestFit analysis of patient 1 (top) and the published sequence (bottom) from fragment 2 of the *annexin* VI gene. The analysis shows no ambiguities between the two sequences suggesting no mutations or polymorphisms exist in this patient.

6.4 Discussion

The *SPARC*, *HAH1* and *annexin* VI genes have been localised and finely mapped to the critical region of the 5q- syndrome at 5q31.3-q32, flanked by the DNA marker D5S413 and the *GLRA1* gene (Boultwood *et al.*, 1994, 2000). Their localisation, expression patterns in $CD34^+$ cells and haematological tissues, and their predicted functions suggest they represent candidates for the putative tumour suppressor gene associated with the development of the 5q- syndrome, and warrant further analysis.

Fine physical mapping of the three candidate genes demonstrated they mapped within YAC 816D6 within 5q32, with *HAH1* mapping immediately adjacent to the *SPARC* gene flanked by the genetic markers D5S1838 and D5S1419 (Boultwood *et al.*, 2000). *Annexin* VI was shown to map proximal to *SPARC* and *HAH1*.

The *SPARC* gene has been shown to regulate cell growth by inhibiting the cell cycle and to possess tumour suppressor activity making it a good candidate. An example of a gene involved in the cell cycle and which possesses tumour suppressor activity is the *p27* (*Kip1*) gene. A major function of *p27* is to bind and inhibit cyclin/cyclin-dependent kinase complexes, thereby blocking cell cycle progression (Philipp-Staheli *et al.*, 2001). The central role of *p27* makes it important in a variety of disease processes that involve aberrations in cellular proliferation and neoplasia. A large number of studies have reported that *p27* expression is frequently downregulated in human tumours. In addition, murine and tissue culture models have shown that *p27* is a potent tumour suppressor gene for multiple epithelially derived neoplasias.

To investigate the proposal that *SPARC* may be mutated in the 5q- syndrome, we directly sequenced three patients with the 5q- syndrome for mutations in the nine coding exons of *SPARC*. A previously unidentified polymorphism in exon 10 was

seen in patient 3 and in normal control DNA. No mutations were found in the nine coding exons of the *SPARC* gene in the three patients with the 5q- syndrome included in the study. Consequently, *SPARC* is unlikely to be the putative tumour suppressor gene associated with the development of the 5q- syndrome.

The *annexin* VI gene has also been shown to possess tumour suppressor activity making it a good candidate for the 5q- syndrome tumour suppressor gene. A member of its family (*annexin* VII) may also play a role in tumour suppression. Other examples of family members both possessing tumour suppressor activity comes from the *p53* family members, *p63* and *p73*. The *p73* and *p63* genes share similarities in transcription activation and apoptosis induction (Tan *et al.*, 2001), and possess 63% amino acid identity in the DNA-binding domain (Levero *et al.*, 2000), with *p53*. Like *p53*, the *p63* and *p73* genes have been found to be mutated in human cancer, although mutations are rare (Irwin and Kaelin, 2001).

We screened the *annexin* VI gene for mutations in the coding region by cycle sequencing. No mutations were found in the coding region of the *annexin* VI gene in nine patients with the 5q- syndrome included in the study. A previously unidentified polymorphism in fragment 4 at nucleotide position 1778 of the coding region, was seen in six out of nine patients and in normal control DNA. *Annexin* VI is unlikely to be the putative tumour suppressor gene associated with the 5q- syndrome.

6.4.1 The role of antioxidants in MDS and leukaemia

Certain dietary (chemical) and endogenous (enzymatic) antioxidants have been cited in the literature as fighting oxidative stress specifically in MDS and leukaemia. A study by Peddie *et al.*, (1997) demonstrated the use of the antioxidant Amifostine in MDS. Amifostine (Ethyol) is an important drug in clinical use which selectively protects normal tissues of various organs from the

effects of radiation and multiple cytotoxic chemotherapeutic drugs (Capizzi and Oster, 2000). Ineffective haematopoiesis in MDS is mediated, at least in part, by apoptosis, though the mechanisms of apoptotic induction are unclear. The authors found that Tumour necrosis factor-alpha (TNF-alpha) promotes apoptosis via intracellular OFR production, oxidation of DNA and proteins, and is increasingly implicated in the pathogenesis of ineffective haematopoiesis in MDS. This data implies a role for intracellular OFR production, mediated by TNF-alpha, in the pathogenesis of ineffective haematopoiesis in MDS, and provides a rationale for the bone marrow stimulatory effects of antioxidants such as Amifostine in MDS.

Arsenic trioxide (As_2O_3) induces remission in a high proportion of patients with acute promyelocytic leukaemia (APL) via induction of apoptosis (Bachleitner-Hofmann *et al.*, 2001). Results suggest that the apoptotic effect of As_2O_3 is not specific for APL but can also be observed in non-APL acute myeloid leukaemia cells. Ascorbic acid has recently been demonstrated to enhance the apoptotic effect of As_2O_3, suggesting a possible future role of As_2O_3/ascorbic acid combination therapy in patients with AML (Bachleitner-Hofmann *et al.*, 2001).

An example of an antioxidant possessing tumour suppressor activity comes from the mitochondrial antioxidant enzyme, manganese-containing superoxide dismutase (MnSOD). Li *et al.*, (2001) showed that reconstitution of MnSOD expression in several human cancer cell lines leads to reversion of malignancy and induces a resistant phenotype to the cytotoxic effects of TNF and hyperthermia, thereby functioning as a tumour suppressor gene.

The *HAH1* gene is thought to play a role in antioxidant defence, suggesting it may be involved in the development of the 5q- syndrome. Eight patients with the 5q- syndrome were sequenced for mutations in the small open reading frame of the *HAH1* gene. No mutations or polymorphisms were found in the 204bp *HAH1*

coding region in the eight patients with the 5q- syndrome included in the study. Consequently, *HAH1* is unlikely to be the gene involved in the pathogenesis of the 5q- syndrome.

6.4.2 Future work

The technique of direct sequencing (cycle sequencing) as the primary method for mutation detection is time-consuming and costly. Since the beginning of this mutation analysis study, the sequencing goals of the HGP have been achieved ahead of schedule. This new sequence data along with the annotation of the critical region of the 5q- syndrome has changed the approach to mutation analysis on candidate genes considered to be associated with the development of the 5q- syndrome.

Future mutation analysis work in this study will be carried out on genomic DNA on the coding exons of each gene mapping to the critical region of gene loss. Each coding exon will be screened by the primary method of DHPLC (denaturing high-performance liquid chromatography). Any sequence changes observed will be confirmed by cycle sequencing. This new approach will save time, money, and valuable patient material.

Chapter 7

Mutation analysis of five 5q- syndrome candidate genes by Denaturing High-Performance Liquid Chromatography (DHPLC)

7.1 Introduction

7.1.1 Identifying mutations in genes implicated in disease

7.1.2 Techniques for mutation detection

7.1.2.1 Single-strand conformation polymorphism (SSCP)

7.1.2.2 Protein truncation test (PTT)

7.1.2.3 Heteroduplex analysis (HDA)

7.1.2.4 Denaturing gradient gel electrophoresis (DGGE)

7.1.2.5 Non-isotopic RNase cleavage assay (NIRCA)

7.1.2.6 Direct sequencing

7.1.2.7 Denaturing high-performance liquid chromatography

7.1.3 The WAVE™ DNA Fragment Analysis System

7.1.4 Mutation analysis on genes mapping to the critical region of the 5q- syndrome, by DHPLC

7.1.4.1 The *GSHPx-3* gene

7.1.4.2 The *MEGF1* gene

7.1.4.3 The *PDGFRβ* gene

7.1.4.4 Novel gene ENSG00000145872

7.1.4.5 Novel gene ENSG00000086589

7.1.4.6 Aims of the study

7.2 Materials and Methods

7.2.1 Candidate gene selection

7.2.2 Ensembl exon prediction

7.2.3 Samples

7.2.4 PCR amplification

7.2.4.1 Exon optimisation

7.2.4.2 High-fidelity PCR

7.2.5 Hybridisation of PCR products to form heteroduplexes

7.2.5.1 Ratio of wild-type to mutant DNA

7.2.6 The WAVE™ DNA Fragment Analysis System

7.2.6.1 WAVEMAKER™ prediction software for mutation detection
7.2.6.2 Loading of samples
7.2.7 Sequencing of heterozygotes
7.2.7.1 Samples
7.2.7.2 Sequencing reactions
7.2.7.3 Data analysis using the Genetics Computer Group (GCG) software package
7.2.8 RACE PCR
7.2.8.1 Data analysis using the Genetics Computer Group (GCG) software package
7.2.9 Genomic PCR

7.3 Results
7.3.1 Candidate gene selection
7.3.2 Ensembl exon prediction
7.3.2.1 The *GSHPx-3* gene
7.3.2.2 The *MEGF1* gene
7.3.2.3 The *PDGFRβ* gene
7.3.2.4 Novel gene ENSG00000145872
7.3.2.5 Novel gene ENSG00000086589
7.3.3 Samples
7.3.4 PCR amplification
7.3.4.1 Exon optimisation
7.3.4.2 High-fidelity PCR
7.3.5 Hybridisation of PCR products to form heteroduplexes
7.3.6 The WAVE™ DNA Fragment Analysis System
7.3.6.1 Temperature prediction
7.3.6.2 Mutation analysis results by DHPLC
7.3.7 Sequencing of heterozygotes
7.3.8 Data analysis using the Genetics Computer Group (GCG) software package
7.3.9 RACE PCR
7.3.10 Data analysis using the Genetics Computer Group (GCG) software package
7.3.10.1 BestFit analysis
7.3.10.2 BlastX analysis
7.3.10.3 Translate
7.3.11 Genomic PCR
7.4 Discussion

7.1 Introduction

7.1.1 Identifying mutations in genes implicated in disease

Over the past two decades there has been a greater understanding of the molecular genetics of human cancer. It is now known that cancer is essentially a genetic disease arising from inherited and/or somatically acquired mutations at different genetic loci, and that tumourigenesis is a multistep process (Pearson and Van der Luijt, 1998). Since the discovery of the cellular basis of hereditary; the chromosome, and the molecular basis of hereditary; the DNA double helix, there has been a quest to decipher first genes and then entire genomes. The plan to determine the complete human genome sequence was established by a consortium in 1995. The first two goals of the Human Genome Project were to; identify all the approximately 100,000 (now believed to be approximately 30,000) protein-coding genes in the human genome; and to determine the sequences of the three billion chemical bases that make up human DNA. This information will help to understand which genes are implicated in disease and to determine which nucleotide(s) changes have functional consequences.

7.1.2 Techniques for mutation detection

There are currently a number of sensitive methods for the detection of changes in the nucleic acid sequence. The introduction of the Polymerase Chain Reaction, which allows specific *in vitro* amplification of a particular target DNA sequence (Saiki *et al.*, 1988), has greatly facilitated the development of techniques to identify genetic alterations.

7.1.2.1 Single-strand conformation polymorphism (SSCP)

Single-strand conformation polymorphism (SSCP) analysis is a rapid method for the detection of minor sequence changes in DNA (Hayashi and Yandell, 1993). Over the last decade, the technique has been widely used to detect mutations in oncogenes, tumour suppressor genes, and genes responsible for genetic diseases (Hayashi and Yandell, 1993). SSCP is particularly useful when searching for small deletions or insertions, single base mutations, and polymorphisms. Large insertions or deletions greater than 1kb are likely to go undetected. Another major disadvantage of the technique is its sensitivity. SSCP is believed to have an 85% detection rate of fragments shorter than 300bp. The sensitivity depends on a number of factors, including the mutation pattern in the target sequence and the temperature. Excessive heat may result in the disappearance of mobility shifts (Glavac and Dean, 1993).

SSCP has been widely used in the mutation analysis of disease genes. HNPCC is one of the most common cancer predisposition syndromes (Yuasa, 2000). Mismatch repair genes, such as *hMSH2* and *hMLH1* have been identified as causative genes for most HNPCC mutations detected by SSCP (Yuasa, 2000). SSCP is more often used as the technique prior to direct sequencing. Mok *et al.*, (1993) used these two techniques to detect mutations in the *p53*, *ras*, and *NF1* genes. SSCP has also been used in conjunction with the protein truncation test (PTT) in the screening of mutations in the *APC* gene. Twenty-nine different mutations in thirty-four cases were identified, that all lead to the formation of premature stop codons (Giarola *et al.*, 1999).

7.1.2.2 Protein truncation test (PTT)

The protein truncation test works by targeting mutations that generate shortened proteins, mainly premature translation termination. The PTT has several

advantages over other mutation detection methods; it has good sensitivity, a low false-positive rate, and can pinpoint the site of the mutation. However, a disadvantage is its use of RNA as the target (Den Dunnen and Van Ommen, 1999).

The PTT has been used widely in the discovery of mutations, particularly in tumour suppressor genes. Tuberous sclerosis (TSC) is an autosomal dominant trait characterised by the widespread development of benign tumours (Sampson and Harris, 1994). LOH has been shown for the regions of chromosomes 9q34 and 16p13 known to harbour TSC genes (Mayer *et al.*, 1999). The authors screened the entire coding regions of the *TSC1* and *TSC2* genes with the PTT. They identified a high proportion of *TSC2* splicing aberrations that strengthens the importance of intronic disease-causing mutations (Mayer *et al.*, 1999). More recently, Wimmer *et al.*, (2000) identified two novel mutations in the *NF1* tumour suppressor gene using the PTT.

7.1.2.3 Heteroduplex analysis (HDA)

Heteroduplex analysis (HDA) is a gel electrophoresis based technique that distinguishes double-stranded heteroduplex molecules that form between a mutant and wild-type DNA strand, from homoduplex molecules. An advantage for the use of this method is that it does not require specialised equipment. The method has been modified in recent years to increase the sensitivity of single-base pair alterations. HDA has been used on its own, or more often, in conjunction with another technique (mainly SSCP) in mutation detection. Like SSCP, HDA has been used to screen for mutations in tumour suppressor genes. HDA was used with temperature gradient gel electrophoresis (TGGE) when screening for mutations in exon 15 of the *TSC1* gene (Hass *et al.*, 2000). Three novel mutations were identified in nine unrelated cases. Furthermore, Hass *et al.*, showed HDA

had a higher sensitivity in detecting frameshift mutations, while TGGE was more sensitive in the detection of base changes. HDA has also been used as the sole technique in the screening of heterozgyous germline mutations in the *RB1* gene of patients with bilateral retinoblastoma (Zhang and Minoda, 1995).

7.1.2.4 Denaturing gradient gel electrophoresis (DGGE)

Denaturing gradient gel electrophoresis (DGGE) allows the rapid screening of single base changes in enzymatically amplified DNA (Fodde and Losekoot, 1994). The technique is based on the migration of double-stranded DNA molecules through polyacrylamide gels containing linearly increasing concentrations of a denaturing agent (urea and formamide). The denaturing gradient can also be generated by temperature; this method is termed temperature gradient gel electrophoresis. DGGE has several advantages; high sensitivity (>95%), improved detection of heterozygotes, and easy isolation of the mutant allele for subsequent sequence determination. Disadvantages of the technique include; the need for special equipment, cost, and being time-consuming.

A number of mutations in genes implicated in disease have been identified by DGGE. Analysis of ten exons of the *APC* gene led to the identification of eight novel germline mutations resulting in frameshifts or stop codons (Fodde *et al.*, 1992). Blanquet *et al.*, (1993) identified germline mutations in the *RB1* gene that also resulted in the generation of stop codons, amino acid substitutions, and alterations in splice sites.

7.1.2.5 Non-isotopic RNase cleavage assay (NIRCA)

NIRCA is an RNase-cleavage-based method for mutation screening that detects mutations as double-stranded cleavage products in duplex RNA targets. This method is reasonably quick, can detect base-pair changes, and the cleaved products can be analysed on agarose gels. The technique is useful for screening large fragments (500bp - 1kb). Smaller fragments (<500bp) may have a mismatch that may not be resolved if cleavage occurs close to one end of the target fragment. As with most other methods NIRCA has been used to screen for mutations in the *p53* gene (Macera *et al.*, 1999). The authors identified two point mutations along with an *Apa*I restriction site polymorphism located in intron 7 within *p53*. The polymorphism allowed the authors to detect LOH in informative samples in a population of patients with prostate cancer. LOH was detected in 10/31 patients (32.4%) suggesting the *p53* tumour suppressor gene may play a more active role in prostate cancer than was previously believed (Macera *et al.*, 1999).

7.1.2.6 Direct sequencing

The most sensitive screening technique for genes that predispose patients for particular cancers is direct sequencing (Gross *et al.*, 1999). However, sequencing of complex genes is technically demanding, costly, and time-consuming. Direct sequencing (cycle sequencing) is often used to confirm and determine the nature of the mutation/polymorphism, following screening by one of the aforementioned methods. *p53* is the most commonly mutated gene in cancers (Bharaj *et al.*, 1998). Several studies have used SSCP followed by cycle sequencing as a method for *p53* gene mutation screening. For example, Wang-Gohrke *et al.*, (1998) identified eleven mutations from forty-four SSCP-negative frozen ovarian cancer samples, and 17/61 (28%) patients with hairy cell leukaemia harboured *p53* mutations in exons 5-8 of the *p53* gene (Konig *et al.*, 2000). Cycle sequencing has also identified mutations in patients with neurofibromatosis type 1 (Luria *et al.*, 1997), and other

tumour suppressor genes, for example *p15* and *p16*, in patients with bladder cancer (Orlow *et al.*, 1995).

7.1.2.7 Denaturing High Performance Liquid Chromatography (DHPLC)

The most recently developed technique for detecting mutations in genes involved in carcinogenesis is DHPLC, also known as Temperature Modulated Heteroduplex Analysis (TMHA). The technique was first pioneered by Transgenomic and named the WAVE™ Nucleic Acid Fragment Analysis System (Transgenomic, Inc., San Jose, CA). DHPLC employs the formation of heteroduplexes between wild-type (reference) and mutated DNA which are efficiently separated on a unique, polymer-based separation matrix with detection by UV/Vis and/or fluorescence. DHPLC has significant advantages over past techniques; experimental time is greatly reduced due to full system automation, parameter prediction and control features of the WAVEMAKER™ utility software; detection of unknown mutations in heterogeneous samples; and fragments are immediately available for direct sequencing, PCR amplification, and cloning. Moreover, there is significant reduction in the number of samples sequenced, and sensitivity has been reported to approach 100%. This non-gel, high-throughput technology provides the ideal platform for cancer research projects.

Genes implicated in cancer and disease, including tumour suppressor genes, are now being screened for mutations by DHPLC. Studies include DHPLC analysis on the *TSC1* gene. DHPLC detected 27/28 (96%) known *TSC1* sequence variations. The only sequence variation not identified was a mosaic case (Roberts *et al.*, 2001). Other studies include blind analysis of exon 16 of the *NF1* gene where 55/55 (100%) individuals were correctly identified (O'Donovan *et al.*, 1998). DHPLC is now being used for new studies, for example, mutation analysis of the entire mitochondria genome has recently been reported (van den Bosch *et al.*, 2000).

DHPLC has been compared with other techniques to clarify its superiority. Gross *et al.*, (1999) conducted a study of *BRCA1* mutation analysis comparing DHPLC with SSCP and direct sequencing. Sequencing is the most sensitive technique, but is time-consuming. SSCP is one of the most frequently used pre-screening methods but its sensitivity and efficiency are unsatisfactory. The DHPLC technique resolved 100% of the DNA alterations that were observed in cycle sequencing. In contrast, mutation analysis by SSCP accounted for 94% of the detected variations. In addition, DHPLC allowed the discrimination between different alterations in a single fragment.

7.1.3 The WAVE™ DNA Fragment Analysis System

The need for automated and high-throughput systems for DNA analysis has increased due to the demands of several genome projects. Many human diseases result from defects in genetic information leading to pathological symptoms and changed phenotypes. Thus, DNA sequence variants (polymorphisms) may be used in the analysis and diagnosis of genetic disease.

The Transgenomic WAVE™ DNA Fragment Analysis System is an accurate, automated, rapid and economical tool to screen for manifestations of changes in DNA sequence and analysis of DNA fragments. Transgenomic's approach for analysis of nucleic acids with the WAVE™ system is based on a liquid chromatography principal.

Analysis on the WAVE™ system is performed at a temperature sufficient to partially denature the DNA heteroduplexes. The melted heteroduplexes are resolved from the corresponding homoduplexes by ion-pair reversed-phase liquid chromatography (DHPLC or TMHA). The differential retention times on the DNAsep® matrix allow for high sensitivity and rapid single nucleotide and short tandem repeat polymorphism (SNP and STR, respectively) detection.

In conclusion, the WAVE™ Fragment Analysis System is the first commercially developed automated technology for DNA sequence variation detection and fragment sizing. It offers considerable advantages for high-throughput screening of SNPs and STRs in the human genome.

7.1.4 Mutation analysis on genes mapping to the critical region of the 5q- syndrome, by DHPLC

We decided to use DHPLC as the primary screening method for mutation screening of coding exons from candidate genes (known and predicted) mapping to the approximate 1.5Mb critical region of the 5q- syndrome at 5q31.3-5q32, flanked by the genetic marker D5S413 and the *GLRA1* gene.

The coding exons for each candidate gene have been predicted using the Ensembl gene prediction program, available from the Sanger Centre (http://www.ensembl.org/). Thirty-six genes (twenty-three known, thirteen predicted (novel)) represented by approximately five hundred coding exons have been predicted to map to the critical region of the 5q- syndrome. This represents a gene-rich region making priority of importance. To coincide with the publication of the draft public Human Genome Sequence on February 15 2001, the Ensembl site has been updated to the October 7 data set, which was the main data set used in the publication and covers 94% of known genes.

We selected genes for analysis that represent good candidate tumour suppressor genes for the 5q- syndrome, i.e. they possess tumour suppressor activity, have antioxidant properties, and have already been implicated in leukaemogenesis. The five candidate genes included in the study were the human plasma glutathione peroxidase (*GSHPx-3*) gene, human homologue of the *Drosophila* tumour suppressor gene *fat2* (*MEGF1*) gene, the human platelet-derived growth

factor receptor, beta (*PDGFRß*) gene, and two novel genes ENSG00000145872, and ENSG00000086589.

7.1.4.1 The *GSHPx-3* gene

GSHPx-3 is one of a family of selenium-dependent glutathione peroxidases that reduce hydrogen peroxide and organic hydroperoxides in the presence of reduced glutathione. The essential role of *GSHPx-3* is its ability to protect haemoglobin from oxidative breakdown in erythrocytes. Many forms of active oxygen such as hydrogen peroxide, lipid hydroperoxides, superoxide, hydroperoxy and hydroxyl radicals, and single oxygen are implicated in human disease (Chu *et al.*, 1992). Evidence exists to support a role for oxidant damage in the pathogenesis of rheumatoid arthritis, cardiovascular disease, immune injury and cancer (Cerutti, 1985; Cross *et al.*, 1987). The antioxidant activity of the glutathione peroxidases, including *GSHPx-3*, may have a protective role in the development of many diseases, including atherosclerosis and carcinogenesis (Halliwell, 1987).

7.1.4.2 The *MEGF1* gene

The *MEGF1* gene is the human homologue of the *Drosophila* tumour suppressor gene *fat2* and has been localised to the critical region of the 5q- syndrome by gene dosage analysis, and sublocalised to YAC 816D6 and BAC clone 17D7 (Fidler *et al*, 2001). Previous studies identified *fat2* as a tumour suppressor gene when two recessive lethal mutations in the *fat2* locus caused hyperplastic, tumour-like overgrowth of larval imaginal discs in *Drosophila* (Mahoney *et al.*, 1991).

7.1.4.3 The *PDGFRβ* gene

The *PDGFRß* gene encodes a cell surface tyrosine kinase receptor for members of the platelet-derived growth factor family. These growth factors are mitogens for

cells of mesenchymal origin (Gronwald *et al.*, 1988). *PDGFRß* has been implicated in the t(5;12)(q33;p13) balanced translocation in a subgroup of patients with CMML (Golub *et al.*, 1994). The consequence of the translocation is expression of a fusion transcript in which the tyrosine kinase domain of *PDGFRß* is coupled to a novel ets-like leukaemia gene, *tel*. The *tel-PDGFRß* fusion demonstrates the oncogenic potential of *PDGFRß* and may provide a paradigm for early events in the pathogenesis of AML (Golub *et al.*, 1994).

7.1.5.4 Novel gene ENSG00000145872

Novel gene 145872 was selected for mutation analysis because it was shown to be expressed in $CD34^+$ cells, and to be the human mitochondrial homologue of the bacterial heat-shock protein (hsp70) co-chaperone, GrpE. Molecular chaperones are defined as proteins that interact with non-native states of other protein molecules (Burston and Clarke, 1995). This activity is important in the folding of newly synthesised polypeptides and the maintenance of proteins in unfolded states suitable for translocation across membranes. The tumour suppressor genes *p53* and *RB1* have been shown to act as chaperones (Lane *et al.*, 1993).

7.1.5.5 Novel gene ENSG00000086589

Novel gene 86589 was selected for mutation analysis because it was shown to be expressed in $CD34^+$ cells and to possess an RNA-binding domain RNP-1 (RNA recognition motif). The RNA recognition motif (RRM) is one of the most common eukaryotic protein motifs. RRM sequences form a conserved globular structure known as the RNA-binding domain (RBD) or the ribonucleoprotein domain. Many proteins that contain RRM sequences bind RNA in a sequence-specific manner (Crowder *et al.*, 2001). A tumour suppressor gene whose protein possesses an RBD is the *WT1* gene (Kennedy *et al.*, 1996).

7.1.6 Aims of the study

The primary aim of the study was to analyse the coding exons of five candidate genes for the 5q- syndrome, for mutations, by DHPLC. DHPLC is an accurate, rapid method for detecting changes in the DNA sequence, and has been used successfully to detect mutations in genes implicated in cancer and disease. The first step was to use the Ensembl program to identify known and novel genes predicted to map to the critical region of the 5q-syndrome. Secondly, Ensembl was used to establish the number of predicted coding exons for each known and novel gene, and to select the genes that were expressed in human bone marrow and $CD34^+$ cells. Expression in $CD34^+$ cells was carried out because MDS is a stem cell disorder. Mutation studies were then carried out on these candidate genes by DHPLC followed by direct sequencing with the aim of identifying the 5q-syndrome gene.

During this study, we found one candidate gene, *MEGF1*, to be downregulated in a number of patients with the 5q- syndrome and AML compared to normal controls. Tumour suppressor genes and growth regulatory genes are frequent targets for methylation defects that can result in aberrant expression. The *p16* gene is one of several tumour suppressor genes that has been shown to be inactivated by DNA methylation in various human cancers (Woodcock *et al.*, 1999). Therefore, the second aim of the study was to establish a methylation map of the promoter region of the *MEGF1* gene and evaluate the methylation status of CpG islands within the promoter region.

7.2 Materials and Methods

7.2.1 Candidate gene selection

Candidate genes for the 5q- syndrome gene were selected based on the following criteria; their localisation to the approximate 1.5Mb critical region of the 5q- syndrome at 5q31.3-q32 flanked by the DNA marker D5S413 and the *GLRA1* gene; their expression in $CD34^+$ cells and haematological tissues; and their predicted function, i.e. has antioxidant properties, or functions as a tumour suppressor gene.

7.2.2 Ensembl exon prediction

The coding exons for each candidate gene were either predetermined and accessible on the GenBank database at NCBI, or predicted by the Ensembl program. The genes were predicted by the Ensembl analysis pipeline from either a Genewise or Genscan prediction followed by confirmation of the exons by comparisons to protein, cDNA and EST databases. Novel genes predicted by Ensembl were confirmed experimentally in our laboratory.

7.2.3 Samples

Fifteen patients with the classical features of the 5q- syndrome, including a 5q deletion as the sole karyotypic abnormality were included in the study. In addition to the 5q- syndrome patients, one patient in transformation to AML, two patients previously with MDS that had transformed to AML, plus five AML patients (for the *MEGF1* gene only), were included in the study. Whole peripheral blood, usually 5mls, was spun down and the plasma removed. Alternatively, granulocyte cells were separated from 40mls of peripheral blood by ficoll gradient centrifugation (Boyum, 1984). The granulocytes showed a high level of purity (≥95%). High molecular weight DNA was obtained from either whole peripheral blood or from the fractionated blood leukocytes by Nucleon® extraction. Whole

peripheral blood DNA and granulocyte DNA fractions from healthy individuals were used as controls. Details of patient samples are shown in Table 7.1.

7.2.4 PCR amplification

The exon-specific primers were designed flanking the coding exons of each gene. The primers were approximately 50% GC and at least 19 bases in length. Additionally, the primers contained either a G or C residue as the last 3'-base, and did not have any regions that could self-anneal or form "hair pin" loops. The primers were dissolved in RNase-free water to a concentration of 100pmol/µl.

7.2.4.1 Exon optimisation

Exon optimisation was carried out using Bio*Taq* DNA polymerase (Bioline UK Ltd., London, UK) (**Chapter 2 section 2.5.2 steps 1-5**).

1. For each 50µl PCR reaction the following were added to a sterile 0.6ml PCR tube;

sterile distilled water	up to 50µl
10x reaction buffer	5µl
50mM $MgCl_2$	1.5µl
dNTP mix	4µl
primer 1	100pmol
primer 2	100pmol
template DNA (≈200ng)	1µl
Taq polymerase (2.0-2.5 units)	0.5µl

2. Details of each primer set are shown in Tables 7.2, 7.3, 7.4 and 7.5. Primers designed to generate a PCR product in the range approximately 150-450bp.

Table 7.1 **Clinical details of 5q- syndrome and AML patients included in the study**

Patient	Sex/age	Cytogenetic karyotype	Sample type
1	F/66	46, XX, del(5)(q31q33)	Peripheral blood DNA
2	F/22	46, XX, del(5)(q31q33)	Granulocyte fraction DNA
3	F/65	46, XX, del(5)(q33-q34)	Granulocyte fraction DNA
4	F/70	46, XX, del(5)(q22-q35)	Granulocyte fraction DNA
5	F/60	46, XX, del(5)(q13-q33)	Peripheral blood DNA
6	F/81	46, XX, del(5)	Peripheral blood DNA
7	M/48	46, XY, del(5)(q13-q33)	Granulocyte fraction DNA
8	M/66	46, XY, del(5)(q13-q31)	Granulocyte fraction DNA
9	F/78	46, XX, del(5)	Granulocyte fraction DNA
10	F/61	46, XX, del(5)(q13-q33)	Granulocyte fraction DNA
11	F/83	46, XX, del(5)(q13-q33)	Granulocyte fraction DNA
12	F	46, XX, del(5)	Granulocyte fraction DNA
13	F	46, XX, del(5)	Peripheral blood DNA
14	F	46, XX, del(5) + myeloma	Peripheral blood DNA
15	F	46, XX, del(5)	Granulocyte fraction DNA
16	F	46, XX, del(5) transforming	Granulocyte fraction DNA
17	F/52	46, XX, del(5)(q13-q33) RAEB → AML	Blast cells DNA
18	M/58	46, XY, del(5)(q15-q35) RAEB → AML	Peripheral blood DNA
19	A	AML	Blast cells DNA
20	B	AML	Blast cells DNA
21	C	AML	Blast cells DNA
22	D	AML	Blast cells DNA
23	E	AML	Blast cells DNA

Table 7.2 Exon primer conditions for the *GSHPx-3* and ENSG00000145872 genes

Gene	Exon	Primer name	Primer sequence 5'-3'	Annealing temperature	PCR Product size
GSHPx-3	1	GPX3Ex1-F2	CAGCCGCCTAGCGATTG	57°C	358bp
		GPX3Ex1-R2	GGGATTGCCCATCTGGC		
	2	GPX3Ex2-F	TTCCTTTCCAGCTCTAACTG	55°C	251bp
		GPX3Ex2-R	TGAAATATGCCATACAGCCC		
	3	GPX3Ex3-F2	AGTAGTTCCAGCGGCAC	55°C	306bp
		GPX3Ex3-R2	CCGATAAATCTCCACCATG		
	4	GPX3Ex4-F2	CACTGCACATTCACTGGC	55°C	361bp
		GPX3Ex4-R2	CAGGTGCCAAGAAATTCC		
	5	GPX3Ex5-F	GGCCTCAAGCAAGGTTGAC	60°C	363bp
		GPX3Ex5-R	CCTCCCCTACATGGTGGAC		
145872	1	145872Ex1-F	GGGTAGCAGATAAAGAGCC	60°C	364bp
		145872Ex1-R	CCATCTGGGAAGATCTCTGG		
	2	145872Ex2-F3	CTCATGTGATCTGACAGCC	62°C	383bp
		145872Ex2-R3	GACTACCTAGCAGTGCTGC		
	3	145872Ex3-F	GTGACTGCGCTCTGGGAG	65°C	458bp
		145872Ex3-R	CGGGCAGAGTGATTTCCTC		

Table 7.3 Exon primer conditions for the *PDGFRß* gene

Gene	Exon	Primer name	Primer sequence 5'-3'	Annealing temperature	PCR Product size
PDGFRß	1	PDGFR-Ex1-F2 PDGFR-Ex1-R2	CTGCCACCAGCACACATC GGCTCATTCTGCAGGAGCC	60°C	220bp
	2	PDGFR-Ex2-F2 PDGFR-Ex2-R2	AGCACTCTCTGGACTTCCC GTGGCCTCCTCGCAGGC	62°C	481bp
	3	PDGFR-Ex3-F PDGFR-Ex3-R	AGAATCCACTGGGAAGTG GGGATGGCCAGAAACCG	57°C	430bp
	4-5	PDGFR-Ex4-5F PDGFR-Ex4-5R	GTATCAAAAATGCAACTC GCTGGTGGTGACTTCCC	57°C	571bp
	6	PDGFR-Ex6-F3 PDGFR-Ex6-R3	TCTAGGAGGGATGAACTGTC ACTCCATGGCTGGCACGG	55°C	477bp
	7	PDGFR-Ex7-F2 PDGFR-Ex7-R2	ACTCCTCCCATGGGTGGG GGGAGAACTGTAAGAGTCC	60°C	266bp
	8	PDGFR-Ex8-F PDGFR-Ex8-R	GGGACTAGATAACCTTCACG AGGACCTGTCCTGTTAACTG	58°C	261bp
	9	PDGFR-Ex9-F PDGFR-Ex9-R	GGTAGGGATTGGGATCGTC AGTTTCCCTGTCTGCAAGG	58°C	316bp
	10	PDGFR-Ex10-F PDGFR-Ex10-R	GCCAGATCACGCAGCATTC ATCTATGATGCCAAAGATGGG	60°C	269bp
	11	PDGFR-Ex11-F PDGFR-Ex11-R	CAGACCTCAGAGAGTCTTC AGACGGACGAACCTAATGG	60°C	275bp
	12-13	PDGFR-Ex12-13F PDGFR-Ex12-13R	CTGGGAGAGGCTAAGTGTG GCAGCTTCCTGGTAGGCC	60°C	228bp
	14	PDGFR-Ex14-F PDGFR-Ex14-R	GTGTGCTGTTGTGCAAGGC AGAATAGGCTCCTGTGGTG	60°C	282bp
	15	PDGFR-Ex15-F PDGFR-Ex15-R	CTCCTCAGGTATCCCAAAG TTGAAGGGACGCCTGAGG	60°C	263bp
	16	PDGFR-Ex16-F PDGFR-Ex16-R	AAGAGCATCAGCCTGTTTGG GAGAAGAAATTCATGAGTGGC	60°C	299bb
	17-18	PDGFR-Ex17-18F PDGFR-Ex17-18R	TCACAGGCACTGTGACTGC CCTGTGGGCCAGAAGGAG	60°C	536bp
	19	PDGFR-Ex19-F PDGFR-Ex19-R	ATGAGTGGTCGAGGTAGAC GAGGTCCTTCCTTGCACTC	60°C	218bp
	20	PDGFR-Ex20-F PDGFR-Ex20-R	AGCCAGTAGAGTTGGATATC CTCTCCTTGTCCTGTAGAAG	60°C	250bp
	21	PDGFR-Ex21-F PDGFR-Ex21-R	CTTGTACTCGGTGTCTGAC CCTTGTTCTGAGAGGCAGC	60°C	269bp
	22	PDGFR-Ex22-F PDGFR-Ex22-R	CTGTGCACAATTTCCTTGGC ATCCCTGAAGGCATTTCTGG	60°C	336bp
	23a	PDGFR-Ex23a-F PDGFR-Ex23a-R	CGAGAGAGACCACAAAGTC CCCCAAGAAGGATGTGAG	60°C	508bp
	23b	PDGFR-Ex23b-F PDGFR-Ex23b-R	CTCACATCCTTCTTGGGG GTATTCCAGGTGGTTGCAC	60°C	449bp
	23c	PDGFR-Ex23c-F PDGFR-Ex23c-R	GTGCAACCACCTGGAATAC CTGGGGCCATTAGGCAGC	60°C	429bp
	23d	PDGFR-Ex23d-F PDGFR-Ex23d-R	GCTGCCTAATGGCCCCAG GAAGAAAACTGCAGGGGCC	62°C	437bp
	23e	PDGFR-Ex23e-F PDGFR-Ex23e-R	GGCCCCTGCAGTTTTCTTG ACTGCTGCTGGAATCCTCC	58°C	440bp

NB. Exons 12 and 13 have now been predicted as one exon. This change does not affect the results in any way.

Table 7.4 Exon primer conditions for 15/23 exons of the *MEGF1* gene

Gene	Exon	Primer name	Primer sequence 5'-3'	Annealing temp	PCR Product size
MEGF1	2	MEG2-F	CCACCATTGTAGAGATCCC	63°C	439bp
		MEG2-R	GGAATGGTGGGTAAGGGTG		
	3	MEG3-F	TATCTTCCTCCCTGAACCC	60°C	239bp
		MEG3-R	TTGCCTCAGTAAAGTGGCC		
	4	MEG4-F	AAGGCCACTAACCAGCATG	65°C	411bp
		MEG4-R	TGCAATTGTCAGCTCAGGTG		
	5	MEG5-F	GAGCTGGTGTATAAGGATGG	60°C	323bp
		MEG5-R	GTCTTCCTGTCTCTTGGCC		
	6	MEG6-F	AAGAAGGCCTTCCATCTCC	60°C	240bp
		MEG6-R	CTAAATCACTGAGGTTGTGG		
	7	MEG7-F	TCAGCATGTCTCCAAGCATG	63°C	398bp
		MEG7-R	GGCTTGCAGTGCACCTTCTC		
	8	MEG8-F	ATCTTGACCCATCCTCTGAG	60°C	318bp
		MEG8-R	CTTACCCTAACCCTGCCTC		
	11	MEG11-F	TGGATCTGAATGCAGTCCC	60°C	276bp
		MEG11-R	CATGGCACTGGGCACTTG		
	14	MEG14-F	GGCCCAACTGCCTCATTG	60°C	296bp
		MEG14-R	GTTCTTGTCCCACAAAGAGC		
	15	MEG15-F	ACCCCACCGCAGGTATC	60°C	234bp
		MEG15-R	ATTCATCTTCTGGACCTGC		
	16	MEG16-F	GCTACCTCATTGCTAACCTC	60°C	250bp
		MEG16-R	GCATCTTCTGCTAGAAGGG		
	17	MEG17-F	GGGACTCATTCTGCTCTTTG	60°C	279bp
		MEG17-R	CCATGGTCACCACCAGAAG		
	18	MEG18-F2	TAGCCCGTTTGATGTCCAG	58°C	438bp
		MEG18-R2	TCCATGTCACAGAGCAGAG		
	19	MEG19-F	GCCTGGGACACCCACATG	63°C	244bp
		MEG19-R	GCTCAATGGGGACCACTTC		
	21	MEG21-F2	CAGAGTACAGAGCGCATTC	63°C	289bp
		MEG21-R2	CAGTCTGCCAATGCCAGG		

NB: The remaining eight *MEGF1* exons were analysed by direct sequencing

Table 7.5 Exon primer conditions for novel gene ENSG00000086589

Gene	Exon	Primer name	Primer sequence 5'-3'	Annealing temperature	PCR Product size
86589	1	86589-Ex1-F2 86589-Ex1-R2	GCGACGTCATGACGCAAAG TATTCCTTCGGCCGCGCC	60°C	302bp
	2	86589-Ex2-F 86589-Ex2-R	CGCCACCCCAAGTGTCAC TACCACAACATCGACAGGAC	60°C	180bp
	3	86589-Ex3-F 86589-Ex3-R	TAGGATGCCAGGGATTTCC CTGAAGGCTAATTTGGTAGAG	60°C	213bp
	4-5	86589-Ex4-5-F2 86589-Ex4-5-R	CCGTGTTATCAGATCTCCAG GGCACCTGCACACTTTCAG	60°C	243bp
	6-7	86589-Ex6-7-F 86589-Ex6-7-R	CCTCCCACTTCTTAGATGAG TGTGTGCATGGTTAGTATGG	60°C	491bp
	8	86589-Ex8-F2 86589-Ex8-R2	CTAACAATAGAGAGACTACG ATATGTAGCTACTCTTGGAAG	58°C	404bp
	9	86589-Ex-9-F2 86589-Ex9-R	GTTCATTTAAGCCTGTTAACC TTCAGTCTCCTGGGTACATC	58°C	333bp
	10-11	86589-Ex10-11-F 86589-Ex10-11-R	TCACAAGGATTTCTCTGGGC TGGGCCAGAGTGCCGAC	60°C	494bp
	12a	86589-Ex12a-F 86589-Ex12a-R	GACCTCGAGTTCAGCTACC GGAATCTGACCATTTTACGTC	60°C	455bp
	12b	86589-Ex12b-F 86589-Ex12b-R	GACGTAAAATGGTCAGATTCC GACTGACAGCCATGCTTTC	60°C	453bp
	12c	86589-Ex12c-F 86589-Ex12c-R	GAAAGCATGGCTGTCAGTC GGAAACTACCAGCCTAACAC	60°C	325bp

NB. Exons 4-5 have now been predicted as one exon. This change does not affect the results in any way.

7.2.4.2 High-fidelity PCR

High-fidelity PCR was carried out using a Hot-start *Taq* polymerase, namely Ampli*Taq* Gold® DNA polymerase (Applied Biosystems, UK), for the *GSHPx-3* gene and Thermo-Start™ DNA polymerase (ABgene®, UK), for the *MEGF1, PDGFRß*, ENSG00000145872 and ENSG00000086589 genes. This enables a clean, single PCR product to be produced. The primers and conditions used were from the exon optimisation experiments.

1a. For each 50µl PCR reaction using Ampli*Taq* Gold® DNA polymerase, the following were added to a sterile 0.6ml PCR tube or 96-well plate;

sterile distilled water	up to 50µl
10x reaction buffer	5µl
25mM $MgCl_2$	3µl
8mM dNTP mix	1.25µl
primer 1	10pmol
primer 2	10pmol
template DNA (≈200ng)	1µl
Taq polymerase (1.25 units)	0.25µl

1b. For each 50µl PCR reaction using Thermo-Start™ DNA polymerase, the following were added to a sterile 0.6ml PCR tube or 96-well plate;

sterile distilled water	up to 50µl
10x reaction buffer	5µl
25mM $MgCl_2$	3µl
2mM dNTP mix	5µl
primer 1	100pmol
primer 2	100pmol
template DNA (≈200ng)	1µl
Taq polymerase (1.25 units)	0.25µl

7.2.5 Hybridisation of PCR products to form heteroduplexes

This step is required to optimise the formation of heteroduplexes and homoduplexes. The patient sample is mixed with a sample of the wild-type DNA, denatured then reannealed under partially denaturing conditions. PCR products were hybridised on the GeneAmp® PCR System 9700 (Applied Biosystems UK) under the following conditions;

Initial denaturation	95°C for 4 minutes
42 cycles of,	95°C minus 1.6°C per cycle for 1 minute
	15 °C hold

7.2.5.1 Ratio of wild-type to mutant DNA

PCR products amplified from patient granulocyte DNA were mixed with PCR products amplified from homozygous wild-type DNA in a 50:50 ratio. This was due to the granulocyte fractions having ≥95% purity. PCR products amplified from patient peripheral blood DNA did not require mixing.

7.2.6 The WAVE™ DNA fragment analysis system

The reagents required for running the WAVE™ system are shown in the Appendix. The data can be rapidly generated due to the TMHA parameter prediction capabilities of the WAVEMAKER™ utility software.

7.2.6.1 WAVEMAKER™ prediction software for mutation detection

The WAVEMAKER™ 4.0 prediction software allows the automated gradient and temperature prediction for discovery of SNP's and other mutations. The Mutation Detection application is used to determine the presence of a mutation in the sample fragment under partially denaturing conditions.

1. Each PCR product from each exon was predicted to have a particular Tm.
2. The temperature(s) required to separate the partially double-stranded DNA were predicted for each exon. For mutation discovery, it is desirable to analyse fragments at several temperatures.
3. Based on the Tm of each fragment, the gradient is calculated for each temperature. The gradient template provides an approach to create a gradient between Buffer A (0.1M TEAA) and Buffer B (0.1M TEAA in 25% acetonitrile). Gradients for mutation detection comprise a DNA loading step, the linear separation gradient, a clean-off step and finally, equilibration.
4. A method is created for each individual temperature required for each exon.
5. A project is constructed combining methods from each exon. The project has to begin with the lowest temperature working up to the highest temperature.

7.2.6.2 Loading of samples

1. For freshly made buffers, the two standards that are used to assess and calibrate the performance of the instrument prior to actual experimental sample analysis are run. They are used to optimise certain parameters, such as temperature, elution gradient, and buffer composition to ensure optimum data

acquisition. The first standard is the DNA Digest Standard. It consists of a restriction digest of plasmodia pUC18 by *Hae*III and represents a pool of 9 DNA fragments with the following sizes (bp): 80,102,174,257,267,298,434, 458, 587. The DNA digest standard is used to assess instrument performance for size-based DNA fragment separations. The second standard is the Mutation Standard. This consists of a combination of two defined 209bp fragments representing A and G alleles at position 168 of the polymorphic DYS271 locus. Upon heating and renaturation, this fragment mixture forms two homoduplexes and two heteroduplexes that are used to check instrument parameters for heteroduplex-based mutation screening.

2. The hybridised samples were aliquoted into a 96-well plate. Usually, 2-10µl of the sample, per temperature, is injected onto the column.

7.2.7 Sequencing of heterozygotes

Patient samples that produce a heteroduplex on analysis should be directly sequenced to identify the sequence change(s). Cycle sequencing reactions were carried out on the ABI PRISM 3100 Genetic analyser (Applied Biosystems) **(Chapter 2 section 2.17)**.

7.2.7.1 Samples

The original PCR product, prior to hybridisation, should be used as the template in the sequencing reactions. If 5µl of the PCR product was loaded on the WAVE, a 1:10 dilution of the PCR product should be made and 1µl of the dilution used as the template in the sequencing reaction. If 10µl of PCR product was loaded, 2µl of the dilution should be used as the template.

7.2.7.2 Sequencing reactions

Sequencing reactions were carried out using both the forward and reverse exon

specific primers to sequence the whole exon, including flanking intronic sequence, to rule out sequencing ambiguities.

7.2.7.3 Data analysis using the Genetics Computer Group (GCG) software package

The position of the sequence change(s) could often be predicted depending on which temperature on the WAVE™ the change(s) was seen. The patient sequence would be compared with the homozygote wild-type sequence and published sequence, using BestFit analysis (**Chapter 3 section 3.2.14.1**) to determine the nature of the sequence alteration.

7.2.8 RACE PCR

The technology of 5' RACE PCR (**Chapter 2 section 2.13**) was used to determine the true 5' end of the *MEGF1* gene. This data would then be used to identify the location of the *MEGF1* promoter, and determine its methylation status. The libraries chosen for the RACE PCR were tissue-specific to the *MEGF1* gene.

1. Gene-Specific Primers were designed from the 5' end of the *MEGF1* cDNA. Details of the primers, including their melting temperature (T_m) are shown in Table 7.6.

Table 7.6 5' RACE PCR primer conditions for candidate gene *MEGF1*

Gene	Primer name	GSP primer sequence 5'-3'	Tm of primer
MEGF1	MEGR3 (GSP1)	CAGAGATGATCCGGTACCTCACTG	64°C
	MEGR2 (GSP2)	AGCTCTCCACATAGGTCTTGGGAG	64°C
	MEGR4 (GSP1)	AATTACCTCCCACTGGATCCCTCC	64°C
	MEGR6 (GSP2)	ACAGCCAAACCATATCAGCCCTGC	64°C
	MEGR7 (GSP1)	ACATGCAGAGGCTGCAGGAAAGC	64°C
	MEGR8 (GSP2)	AGCCCTGCCTCTGTCATTTGCTAG	64°C
	MEGR10 (GSP1)	GTGGCATCAGGCTGCCTGGCTG	64°C
	MEGR11 (GSP2)	GTAGAATAAGGATAAGAAAATCATCAATC	64°C

2. The Marathon-Ready™ cDNA templates used in the 25µl RACE PCR reaction included; human pituitary gland, skeletal muscle, testis, whole brain, and placenta.
3. The RACE PCR products were purified with Wizard preps as previously described (**Chapter 2 section 2.5.1.3, steps 2-13**), and prepared for sequencing as previously described (**Chapter 2 section 2.19**).

7.2.8.1 Database analysis using the Genetics Computer Group (GCG) software package

The sequence generated from each RACE PCR product was first compared with the *MEGF1* cDNA sequence, using BestFit analysis (**Chapter 3 section 3.2.14.1**) to determine the homology of the overlap. The sequence was then subjected to a BLAST protein search as previously described (**Chapter 3 section 3.2.12.2**) utilising the Mammalian sequences database and the Genome sequences (gss and htg) database. A translation was carried out on the complete RACE sequence as previously described (**Chapter 3 section 3.2.12.4**) to determine if the sequence was in-frame or untranslated.

7.2.9 Genomic PCR

A genomic PCR (**Chapter 2 section 2.5.2**) was carried out for the *MEGF1* gene to confirm the genomic sequence order of the contig that included the *MEGF1* gene, was correct according to the GenBank database. This was achieved by amplifying across the genomic sequence from the RACE sequence using the reverse RACE gene specific primer

1. PCR primers were designed from the part of the contig containing the *MEGF1* gene that was approximately 1kb upstream of the 5'end of the true cDNA (as determined by RACE PCR (**section 7.2.8**). The approximately 1kb fragment was split into two overlapping PCR products using genomic sequence-specific primers and the 5' RACE gene-specific primer. Details of the primers are shown in Table 7.7.

Table 7.7 Genomic PCR primer conditions for candidate gene *MEGF1*

Gene	Primer name	GSP primer sequence 5'-3'	Tm of primer	PCR size
MEGF1	MEG-F1	ATGGGCTCTGTGGGAAACAGCAAG	64°C	580bp
	MEGR9	CCTCATGAGCCTTCATTTCACCTTC	63°C	
	MEG-F2	GAAGAATCTGCCACCTTCCTGCC	64°C	547bp
	MEGR12	GTAAGAACTAGTCCCTGGGAGTTG	63°C	

2. The PCR products were purified with Wizard preps as previously described (**Chapter 2 section 2.5.1.3, steps 2-13**), and prepared for sequencing as previously described (**Chapter 2 section 2.19**).
3. The sequence generated from the PCR products was compared with the genomic sequence from the contig containing the *MEGF1* gene using BestFit analysis as previously described (**Chapter 3 section 3.2.14.1**).

7.3 Results

7.3.1 Candidate gene selection

Five candidate tumour suppressor genes; human plasma glutathione peroxidase-encoding (*GSHPx-3*) gene, human homologue of the *Drosophila* tumour suppressor gene *fat2* (*MEGF1*) gene, human platelet-derived growth factor receptor, beta (*PDGFRß*) gene, and novel genes ENSG00000145872 and ENSG00000086589, were selected for mutation analysis by DHPLC. The known candidate genes had previously been mapped to the approximate 1.5Mb critical region of the 5q- syndrome at 5q31.3-q32 by gene dosage analysis. All candidate genes had previously been shown to be expressed in haematological tissues including human bone marrow, and $CD34^+$ cells. Moreover, all five genes had putative functions that made them candidates for the 5q- syndrome gene.

7.3.2 Ensembl exon prediction

7.3.2.1 The *GSHPx-3* gene

The genomic structure of the *GSHPx-3* gene had previously been determined (Yoshimura *et al.*, 1994). Therefore, sequence data from the 5 exons of *GSHPx-3* was accessed from GenBank at NCBI under the accession numbers: D16360 (exon 1); D16361 (exon 2); and D16362 (exons 3, 4, and 5). The exons had a combined length of 681bp.

7.3.2.2 The *MEGF1* gene

The Ensembl program had predicted the *MEGF1* gene to contain twenty-three coding exons from its 14536bp cDNA. Fifteen of these 23 exons were analysed for mutations by DHPLC. The 8 *MEGF1* exons not analysed by DHPLC were analysed by direct sequencing as the size of these exons were greater than 450bp.

7.3.2.3 The *PDGFRß* gene

The full cDNA sequence of the *PDGFRß* gene was known and was accessed from GenBank at NCBI under the accession number NM_002609. The Ensembl program predicted 23 coding exons comprising a total length of 5216bp. The latest Ensembl prediction has predicted exons 12 and 13 to be one exon, making 22 *PDGFRß* coding exons. This has not affected the results in any way.

7.3.2.4 Novel gene ENSG00000145872

The Ensembl program predicted novel cDNA 145872 to have 3 coding exons with a total length of 600bp.

7.3.2.5 Novel gene ENSG00000086589

The Ensembl program predicted novel cDNA 86589 to have 12 coding exons with a total length of 2300bp. The latest Ensembl prediction has predicted exons 4 and 5 to be one exon, making eleven 86589 coding exons. This has not affected the results in any way.

7.3.3 Samples

Ten patients with the 5q- syndrome were selected from the pool of patients used in the study, see Table 7.1. In addition, the two MDS cases that had transformed to AML were included in the analysis for each gene. Five AML cases with a 5q deletion were included in the mutation analysis of the *MEGF1* gene.

7.3.4 PCR amplification

The PCR conditions for each coding exon from the 5 candidate genes were optimised with Bio*Taq* DNA polymerase. A more specific PCR product was obtained using one of the high-fidelity *Taq* polymerases.

7.3.4.1 Exon optimisation

Each coding exon from the 5 candidate genes was successfully optimised using Bio*Taq* DNA polymerase, see Figure 7.1.

7.3.4.2 High-fidelity PCR

Each coding exon from the five candidate genes was successfully amplified with Ampli*Taq* Gold® DNA polymerase or ThermoStart™ DNA polymerase, see Figure 7.2.

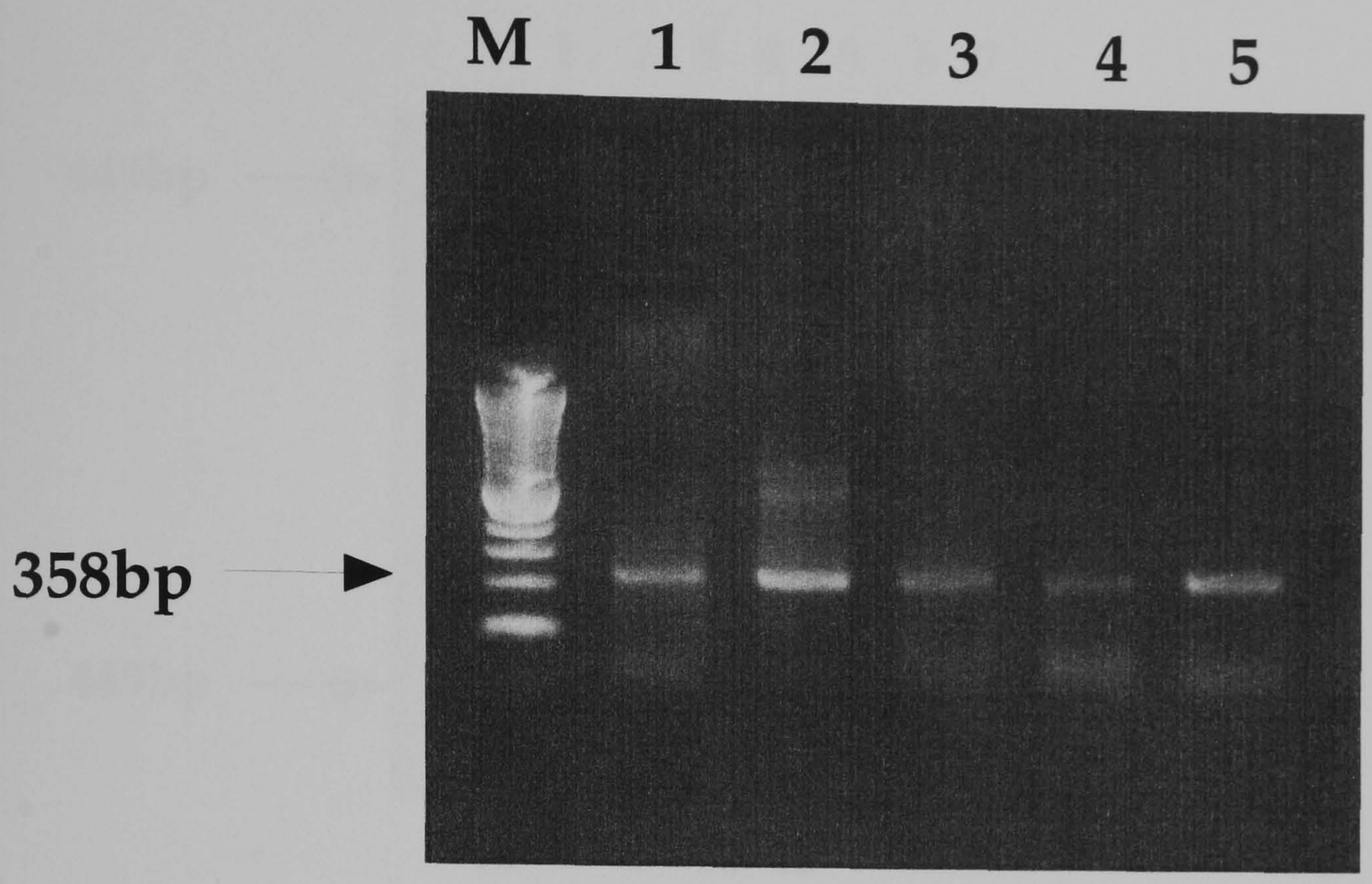


Figure 7.1

Representative agarose gel analysis of the 358bp product of *GSHPx-3* exon 1. Genomic DNA obtained from normal healthy controls (tracks **1** to **5**) was amplified with exon-specific primers, and PCR performed on a thermal cycler. The PCR products were sized with the HyperLadder IV DNA marker (**M**).

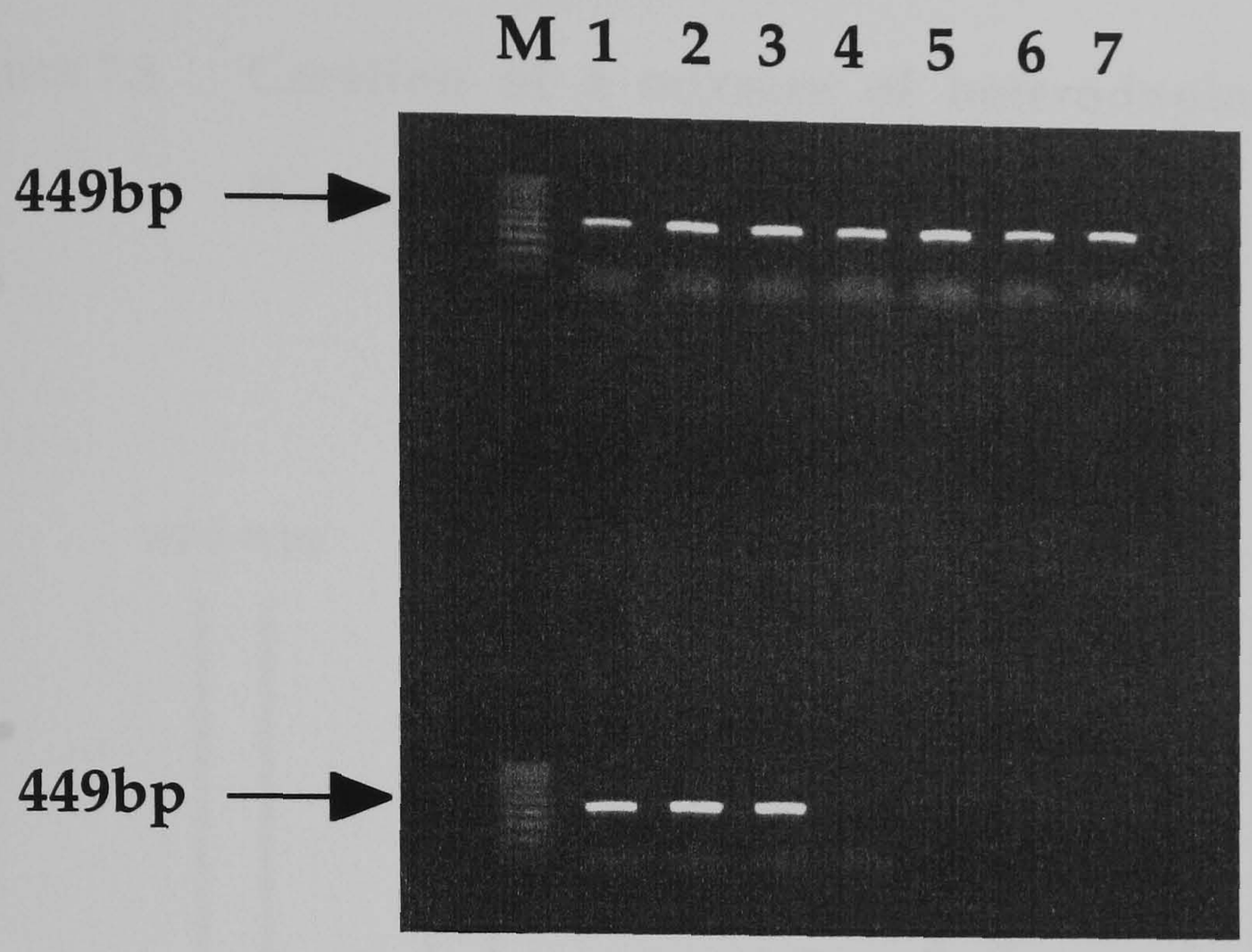


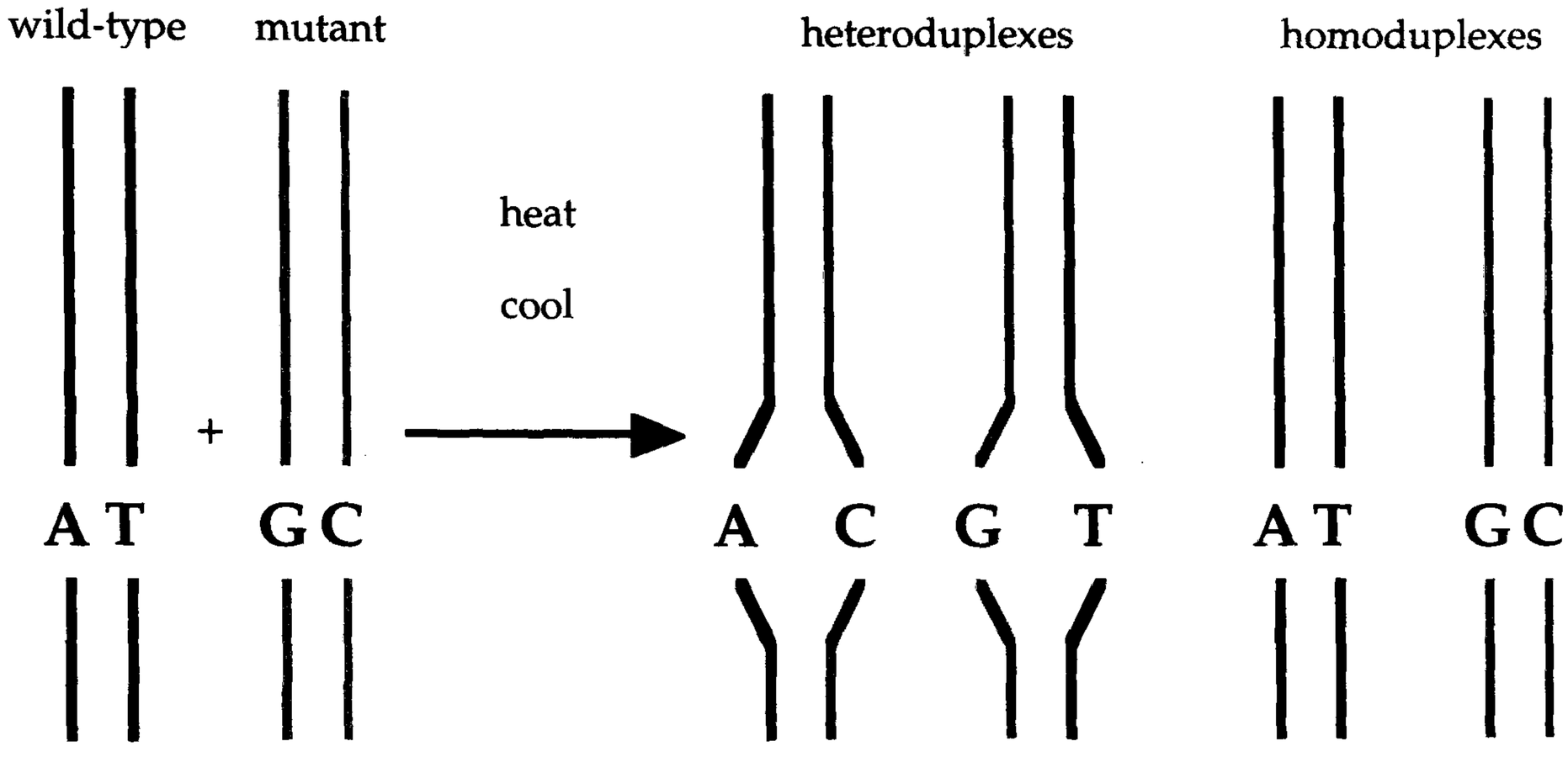
Figure 7.2

Representative agarose gel analysis of the 449bp PCR product of *PDGFRß* exon 23b. Genomic DNA obtained from normal healthy controls (lanes **1** to **10**) was amplified with exon-specific primers, and PCR performed on a thermal cycler. A negative control (lane **11**) was carried out alongside. The PCR products were sized with the SuperMid DNA marker (**M**).

7.3.5 Hybridisation of PCR products to form heteroduplexes

Each PCR product was hybridised to form heteroduplexes, see Figure 7.3.

Figure 7.3 Creation of a mixture of heteroduplexes and homoduplexes through hybridisation



7.3.6 The WAVE™ DNA fragment analysis system

The amplified PCR products from patient and control DNA for each coding exon from each candidate gene were optimised for DHPLC analysis.

7.3.6.1 Temperature prediction

The nucleotide sequence of each coding exon from the 5 candidate genes was analysed using the WAVEMAKER™ utility software. The T_m of the sequence was calculated and a melting curve obtained, see Figure 7.4a.

Figure 7.4a

Graphical representation of the melting curve of *MEGF1* exon 19. The 243bp sequence has a Tm of 62°C and is 53% GC rich.

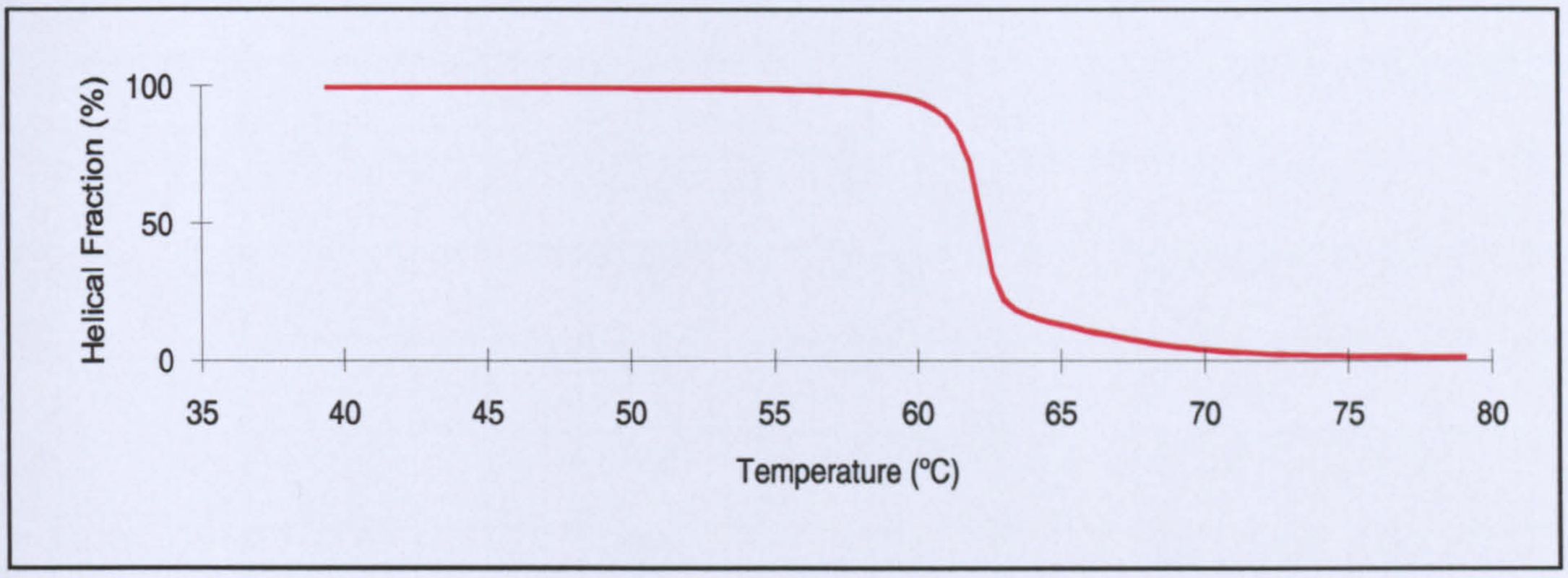
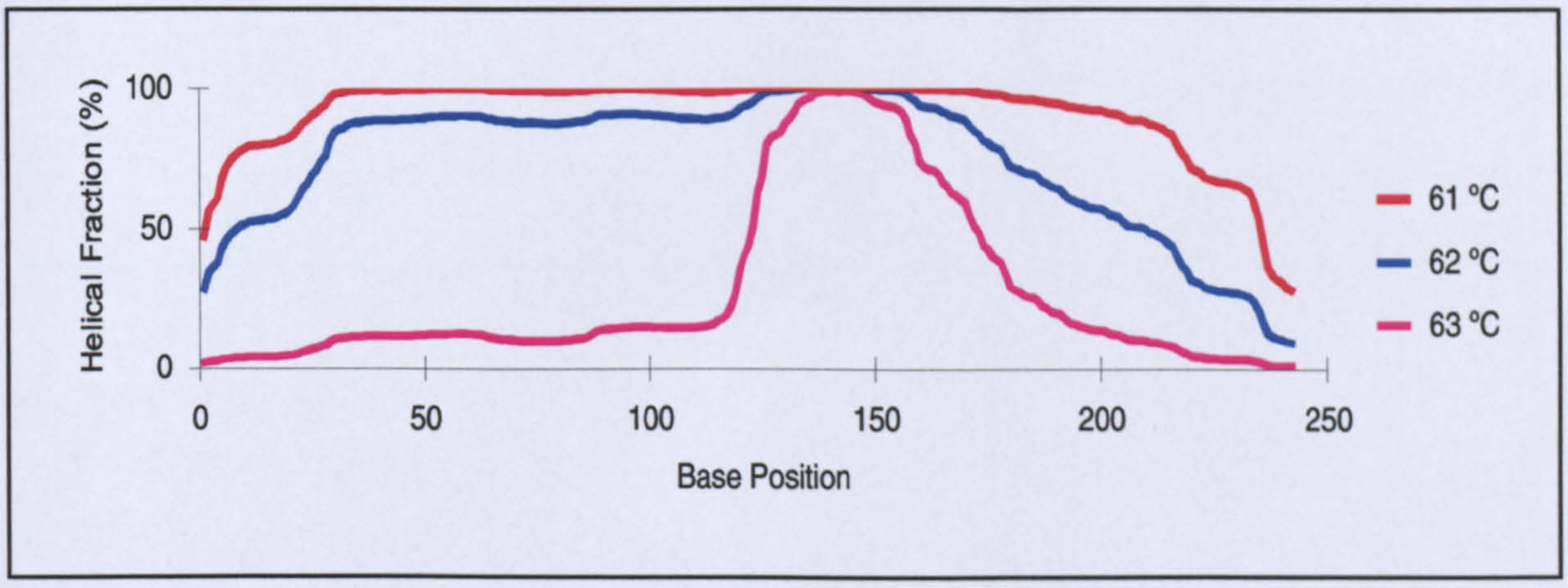


Figure 7.4b

Graphical representation of the temperatures required to partially denature *MEGF1* exon 19. A polymorphism/mutation should only be detected if the DNA retains ≥75% helicicity. Three temperatures are required to cover the 243bp size fragment.



The temperature(s) required to partially denature the DNA of each fragment was then calculated using the WAVEMAKER™ utility software, see Figure 7.4b. The temperatures required for each exon from each candidate gene are shown in Tables 7.8, 7.9 and 7.10.

Table 7.8 **Temperature predictions for DHPLC analysis on the WAVE™ system, for candidate gene MEGF1**

Gene	Exon	PCR product size	Tm of PCR fragment	Temperatures predicted for mutation analysis
MEGF1	2	439bp	58°C	58°C/60°C/61°C/62°C
	3	239bp	61°C	61°C/63°C
	4	411bp	62°C	61°C/62°C/63°C
	5	323bp	63°C	61°C/63°C/64°C
	6	240bp	62°C	61°C/62°C/63°C
	7	398bp	61°C	61°C/62°C/63°C
	8	318bp	61°C	61°C/62°C/64°C
	11	276bp	59°C	59°C/60°C/61°C
	14	296bp	62°C	61°C/62°C/63°C
	15	234bp	61°C	61°C/62°C/64°C
	16	250bp	60°C	59°C/62°C
	17	279bp	62°C	62°C/63°C/64°C
	18	438bp	62°C	61°C/63°C/64°C/65°C
	19	244bp	62°C	61°C/62°C/63°C
	21	289bp	62°C	62°C

Table 7.9 Temperature predictions for DHPLC analysis on the WAVE™ system, for candidate genes PDGFRß and ENSG00000086589

Gene	Exon	PCR product size	Tm of PCR fragment	Temperatures predicted for mutation analysis
PDGFRß	1	220bp	65°C	65°C/66°C
	2	481bp	63°C	62°C/63°C/64°C
	3	430bp	61°C	61°C/62°C
	4-5	571bp	62°C	61°C/62°C/64°C
	6	477bp	65°C	65°C/66°C
	7	266bp	64°C	64°C/65°C
	8	261bp	64°C	64°C/65°C/66°C
	9	316bp	66°C	66°C/68°C
	10	269bp	61°C	62°C/63°C/64°C
	11	275bp	62°C	63°C/65°C
	12-13	228bp	63°C	63°C/64°C/66°C
	14	282bp	65°C	65°C/66°C
	15	263bp	63°C	63°C/64°C/66°C
	16	299bp	61°C	61°C/63°C/64°C
	17-18	536bp	63°C	62°C/63°C
	19	218bp	63°C	62°C/64°C
	20	250bp	61°C	61°C/63°C/64°C
	21	269bp	63°C	63°C
	22	336bp	63°C	63°C/64°C/65°C
	23a	508bp	55°C	55°C/57°C/61°C/64°C
	23b	449bp	60°C	57°C/59°C/62°C/63°C/64°C
	23c	429bp	60°C	57°C/59°C/62°C/64°C
	23d	437bp	61°C	60°C/62°C/63°C
	23e	440bp	63°C	62°C/63°C/65°C/66°C
86589	1	302bp	66°C	65°C/66°C
	2	180bp	61°C	61°C/62°C
	3	213bp	55°C	55°C/56°C
	4-5	243bp	60°C	59°C/61°C/63°C
	6-7	491bp	57°C	55°C/57°C/60°C/61°C
	8	404bp	55°C	56°C/57°C/59°C
	9	333bp	55°C	55°C/58°C/62°C
	10-11	494bp	60°C	60°C/61°C/62°C
	12a	455bp	55°C	54°C/55°C/57°C/60°C
	12b	453bp	56°C	55°C/57°C/58°C/60°C
	12c	325bp	59°C	57°C/58°C/62°C/63°C

NB. *PDGFRß* exons 12 and 13 have now been predicted as one exon. 86589 exons 4 and 5 have now been predicted as one exon. These changes do not affect the results in any way.

Table 7.10 Temperature predictions for DHPLC analysis on the WAVE™ system, for candidate genes *GSHPx-3* and ENSG00000145872

Gene	Exon	PCR product size	Tm of PCR fragment	Temperatures predicted for mutation analysis
GSHPx-3	1	358bp	65°C	63°C/65°C/67°C
	2	251bp	60°C	58°C/62°C
	3	306bp	61°C	56°C/61°C/62°C
	4	361bp	61°C	60°C/61°C/66°C
	5	363bp	60°C	60°C/63°C
145872	1	364bp	54°C	55°C/59°C/63°C
	2	383bp	56°C	56°C/57°C/58°C
	3	458bp	60°C	59°C/60°C/62°C

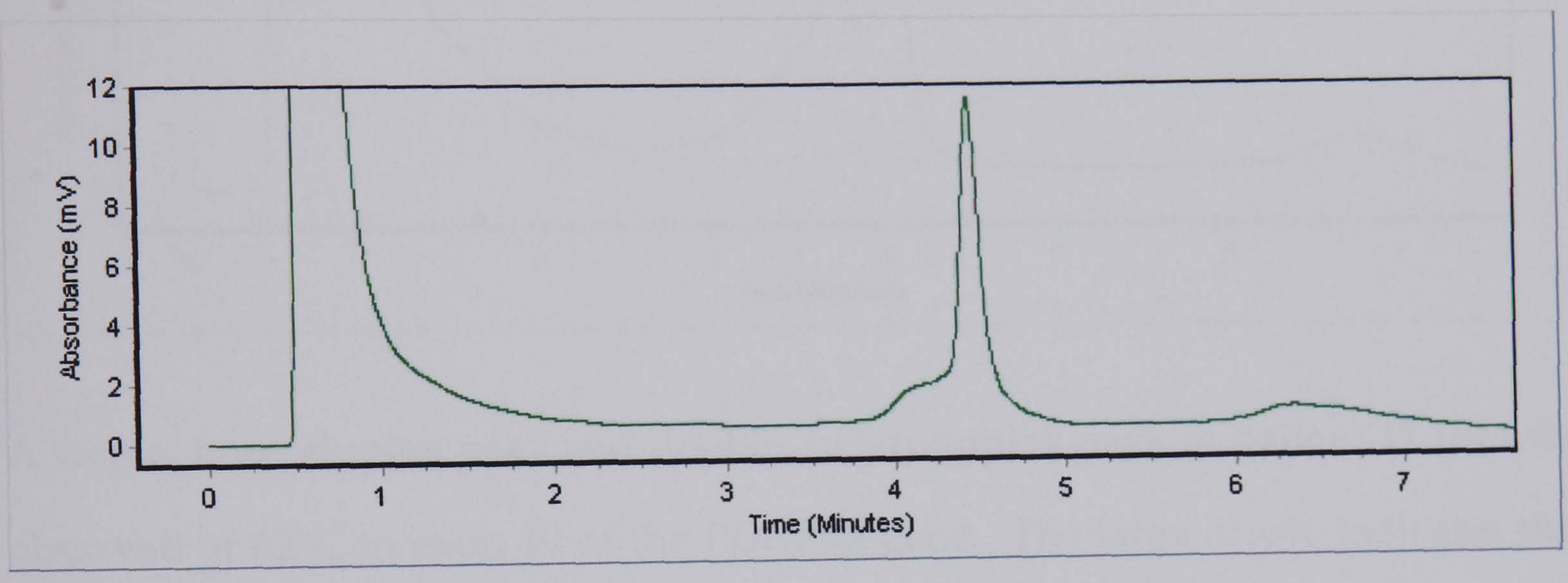
7.3.6.2 Mutation analysis results by DHPLC

The 5q- syndrome and AML patients, plus the normal controls were shown to be homozygous or heterozygous for each coding exon of the candidate genes, by DHPLC. A single peak on the WAVE™ chromatogram represented the patients and controls that were homozygous for a particular exon, see Figure 7.5a. Two to four peaks represented the heterozygous patients and controls, see Figure 7.5b.

Figure 7.5a

Representative DHPLC chromatogram of exon 1 of the ENSG00000145872 novel gene. A single, homozygous peak in patient 3 (a) was observed at 59°C (one of the temperatures predicted by the WAVEMAKER™ software). The amplified PCR product from the patient DNA was mixed with the amplified PCR product from the homozygous wild-type DNA in a 50:50 ratio. This single, homozygous peak was also observed in the normal control (b).

a)



b)

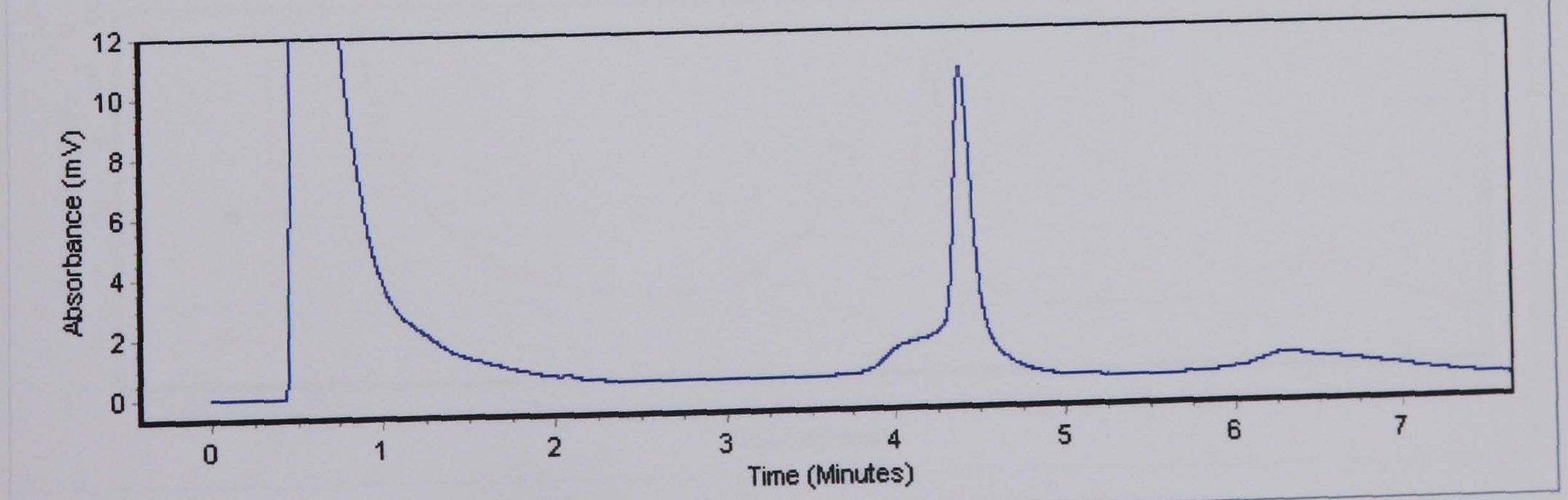


Figure 7.5b

Representative DHPLC chromatograms of heterozygotes. A single, homoduplex and single, heteroduplex peak in patient 5 (a) was observed at 64°C in exon 22 of the *PDGFRß* gene. The large arrow indicates the homoduplex peak. The small arrow indicates the heteroduplex peak.

a)

A single, homoduplex peak and double, heteroduplex peak in patient 11 (b) was observed at 62°C in exon 19 of the *PDGFRß* gene. The large arrow indicates the homoduplex peak. The small arrows indicate the two heteroduplex peaks.

b)

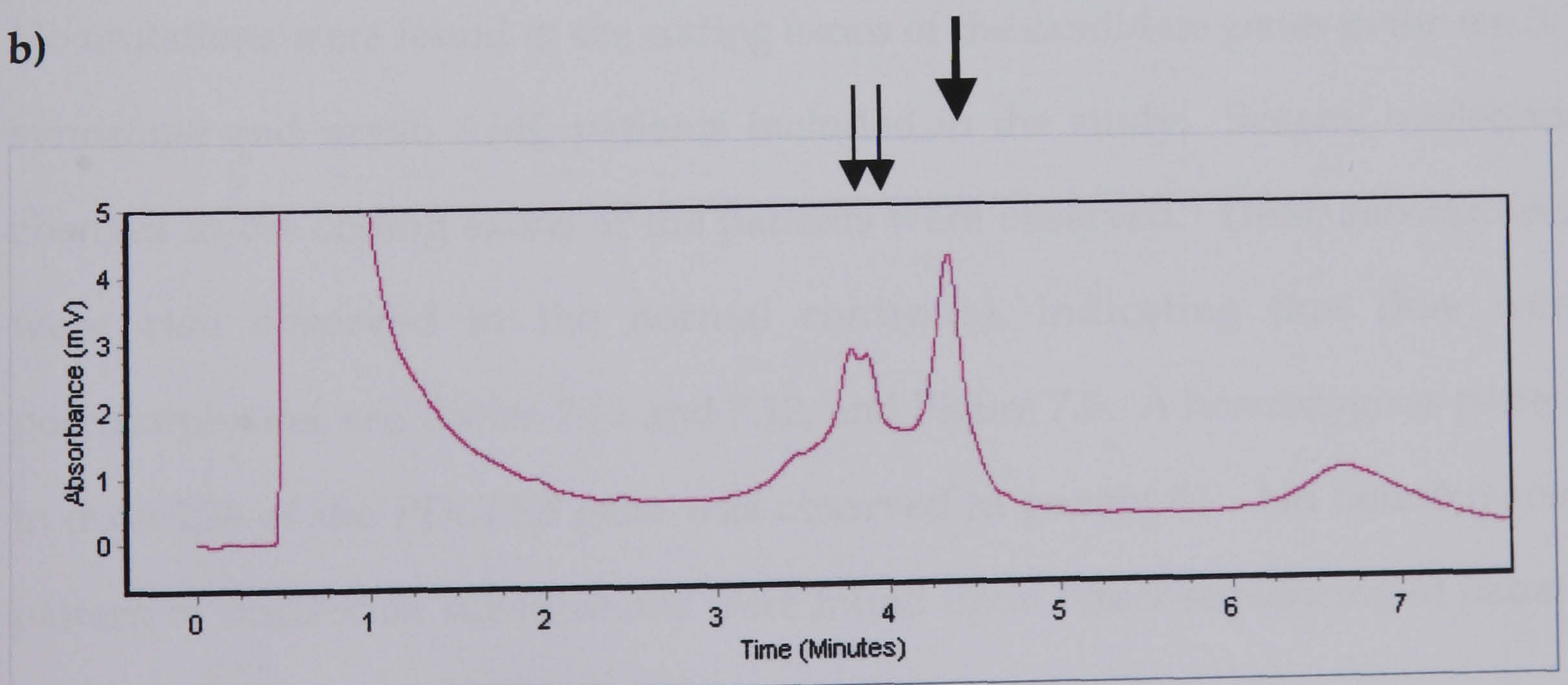
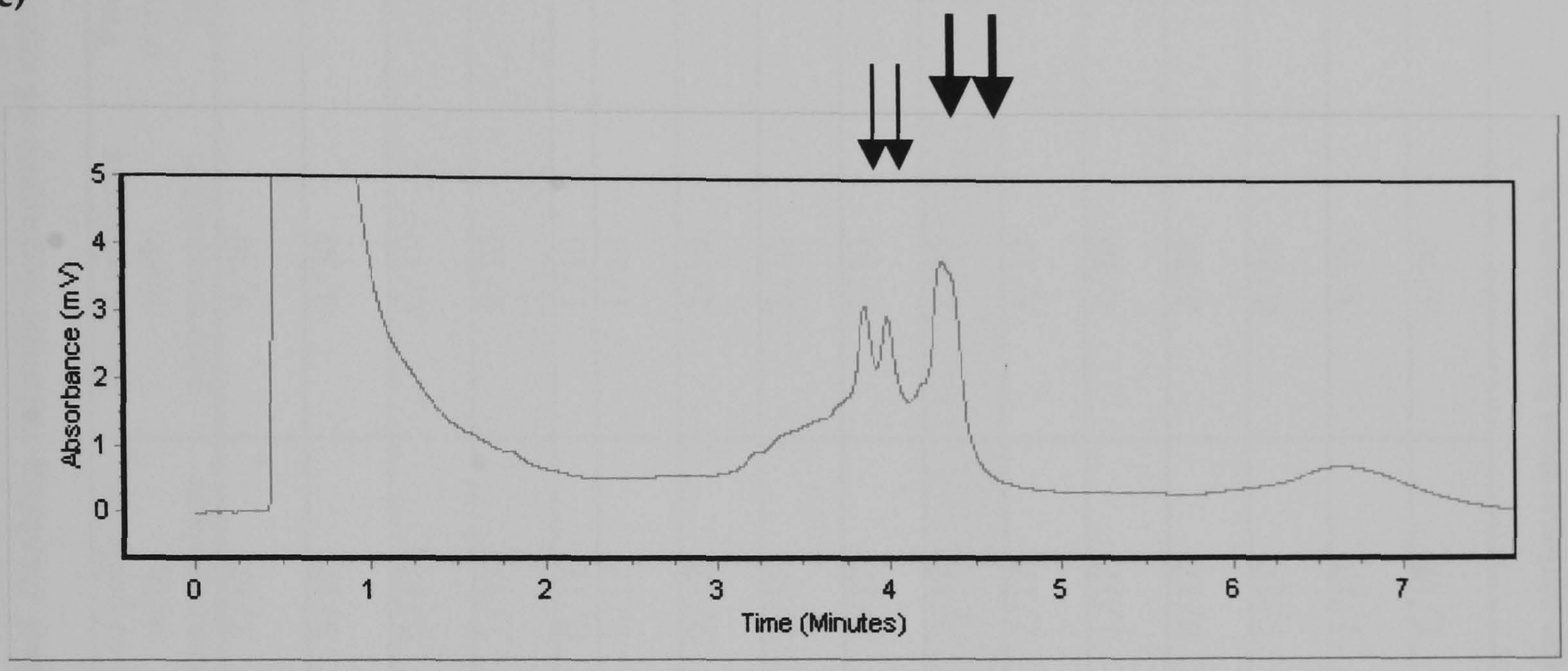


Figure 7.5b

A double, homoduplex and double, heteroduplex peak in patient 12 (c) was observed at 61°C in exon 2 of the *MEGF1* gene. The amplified PCR products from the patient DNA was mixed with the amplified PCR products from the homozygous wild-type DNA in a 50:50 ratio. The large arrows indicate the homoduplex peaks. The small arrows indicate the heteroduplex peaks.

c)



7.3.7 Sequencing of heterozygotes

No mutations were found in the coding exons of the candidate genes in the ten 5q- syndrome and seven AML patients included in the study. Several nucleotide changes in the coding exons of the patients were observed. These substitutions were also observed in the normal control(s), indicating that they were polymorphisms, see Tables 7.11 and 7.12, and Figure 7.6. A heterozygous pattern in exon 23e of the *PDGFRß* gene was observed in patient 11. No heterozygous pattern or nucleotide substitutions were found upon direct sequencing of patient 11, i.e .the patient was homozygous.

Table 7.11 **Frequency of polymorphisms identified from the coding exons and flanking intronic sequence of candidate gene *MEGF1*, by DHPLC, in patients with the 5q- syndrome/AML and normal controls**

Gene	No. of coding exons	No. of polymorphisms identified	Location and sequence change	No. of patients with polymorphism	No. of controls with polymorphism	Frequency of polymorphism in patients	Frequency of polymorphism in controls
MEGF1	23	16	Exon 2 @ 204110 A/G Heterozygote	3/17	9/20	18%	45%
			Exon 2 @ 204226 A/G Heterozygote	3/17	9/20	18%	45%
			Intron flanking Exon 3 A/T Heterozygote	3/17	0/10	18%	—
			Exon 5 C/T Heterozygote*	1/17	0/20	6%	—
			Exon 6	0/17	1/10	—	10%
			Exon 7 @ 23623 A/G Heterozygote	0/17	1/10	—	10%
			Exon 7 @ 23643 C/T Heterozygote	2/17	2/10	12%	20%
			Exon 7 @ 23802 C/T Heterozygote	2/17	2/10	12%	20%
			Exon 7 @ 23831 C/G Heterozygote	0/17	1/10	—	10%
			Exon 8 C/T Heterozygote	3/17	2/45	18%	4%
			Exon 14 @ 44226 A/G Heterozygote	2/17	3/10	12%	30%
			Exon 14 @ 44227 A/G Heterozygote	1/17	1/10	6%	10%
			Exon 17 @ 47554 A/G Heterozygote	0/17	2/26	—	8%
			Exon 17 @ 47641 A/G Heterozygote	2/17	8/26	12%	31%
			Exon 18 @ 69362 C/T Heterozygote	1/17	3/10	8%	30%
			Exon 18 @ 69401 A/G Heterozygote	1/17	3/10	8%	30%

* This nucleotide change did not change the amino acid at this position
The highlighted boxes show where the frequency of the polymorphism is higher in the patients than in the controls

Table 7.12 **Frequency of polymorphisms identified from the coding exons and flanking intronic sequence of candidate genes GSHPx-3, PDGFRß, and ENSG00000086589 by DHPLC, in patients with the 5q- syndrome/AML and normal controls**

Gene	No. of coding exons	No. of polymorphisms identified	Location and sequence change	No. of patients with polymorphism	No. of controls with polymorphism	Frequency of polymorphism in patients	Frequency of polymorphism in controls
GSHPx-3	5	1	Exon 5	1/10	1/10	10%	10%
86589	11 (12)	5	Intron flanking Exon 8 A/G Heterozygote	1/12	0/20	8%	—
			Intron between Ex10-11 A/G Heterozygote	4/12	23/48	33%	48%
			Exon 10	1/12	1/48	8%	2%
			Exon 12	0/12	1/10	—	10%
PDGFRß	22 (23)	16	Intron flanking Exon 2	1/12	1/10	8%	10%
			Exon 3	0/12	1/10	—	10%
			Ex 6 A/G Heterozygote	1/12	1/20	8%	5%
			Exon 7	0/12	1/10	—	10%
			Intron flanking Exon 9	2/12	2/10	17%	20%
			Exon 12-13	0/12	1/10	—	10%
			Intron flanking Exon 16 C/T Heterozygote	1/12	1/10	8%	10%
			Exon 19 C/T Heterozygote	6/12	6/10	50%	60%
			Intron flanking exon 22	5/12	5/10	42%	50%
			Exon 22 A/G Heterozygote	2/12	1/10	17%	10%
			Exon 23a	0/12	1/10	—	10%
			Exon 23c A/G Heterozygote	1/12	1/45	8%	2%
			Exon 23d				
			Exon 23e @ 72432	3/12	3/10	25%	30%
			Exon 23e @ 72466	1/12	1/10	8%	10%
			Exon 23e @ 72574	4/12	4/10	33%	40%

The highlighted boxes show where the frequency of the polymorphism is higher in the patients than in the controls

a)

b)

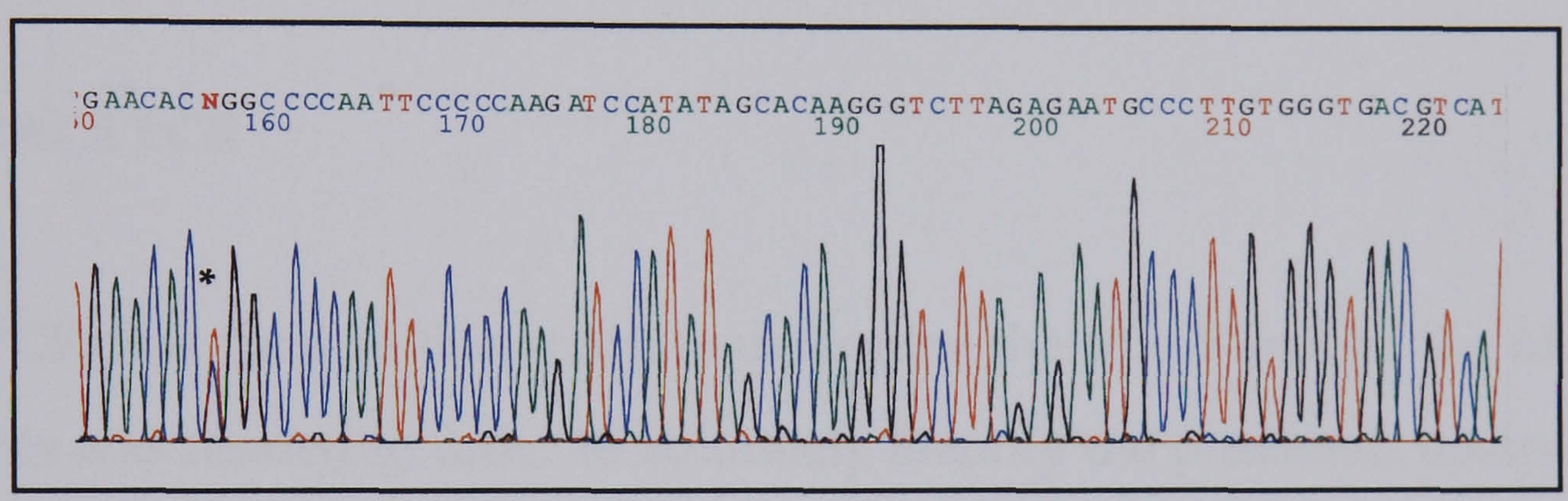


Figure 7.6

Representative sequence analysis of exon 14 of the *MEGF1* gene. A heterozygous pattern of C and T alleles in patient 21 (a) was observed at nucleotide 44226. This heterozygous pattern was also observed in the normal control (b). An asterisk (*) indicates the position of the heterozygote.

Direct sequencing of the wild-type DNA that had been mixed with patient 11 showed a G to A substitution at nucleotide 72432 when compared with the published sequence. Patient 11 has a G at nucleotide 72432. Therefore, patient 11 and the normal control were homozygous for alternate alleles at nucleotide 72432, see Figures 7.7a and 7.7b.

7.3.8 Data analysis using the Genetics Computer Group (GCG) software package

Sequence data from each patient was compared with the published sequence and the normal control using BestFit analysis, see Figure 7.8.

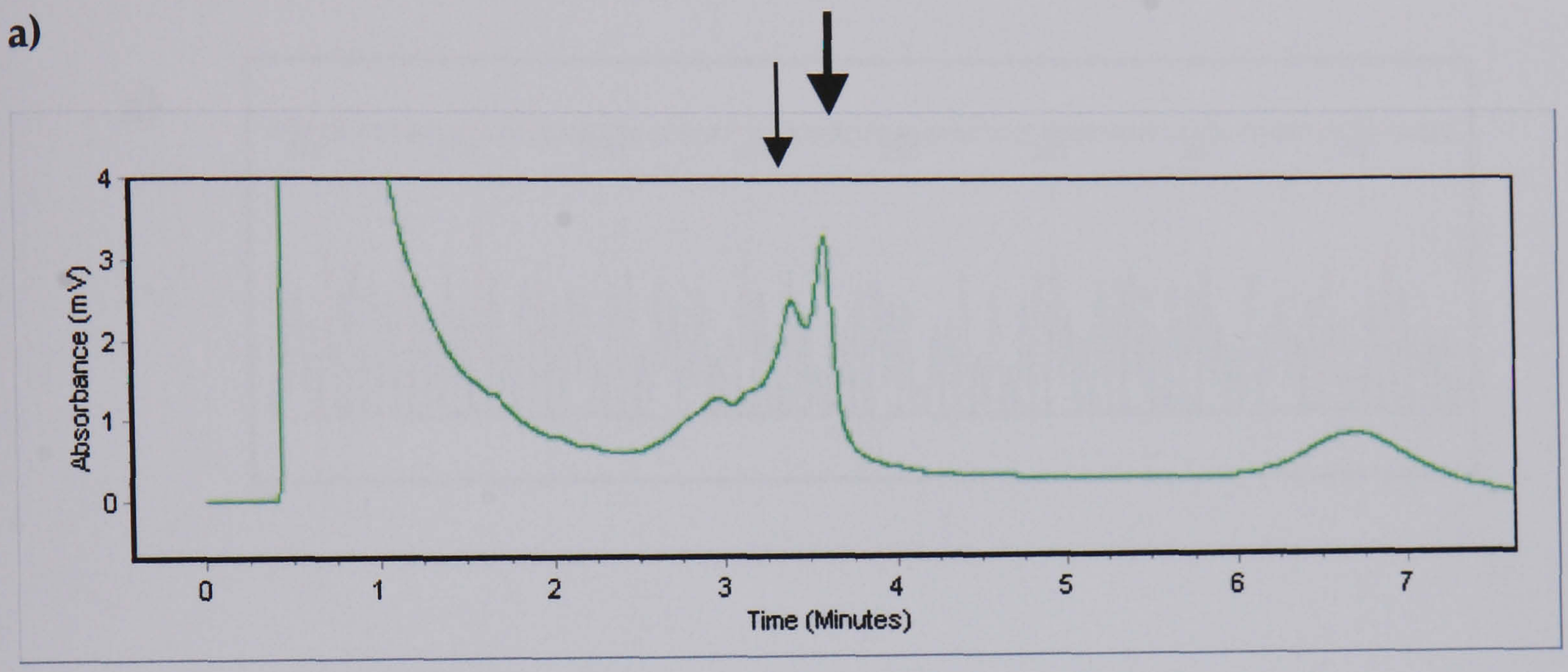
7.3.9 RACE PCR

RACE PCR was used in this study to determine the true 5' end of the *MEGF1* gene. This was needed in order to accurately identify the promoter, upstream of the 5' end. Three 5' RACE PCR products were generated from the human pituitary gland, skeletal muscle, testis, and whole brain cDNA libraries with gene-specific primers designed from the 5' sequence of the *MEGF1* cDNA. Direct sequencing of the first RACE PCR reaction generated products of various sizes. The largest product (865bp) was generated from the skeletal muscle cDNA library, and selected for further analysis. The second (501bp) and third (~240bp) RACE PCR reactions generated products of the same size in all tissues. Direct sequencing of the first two RACE products generated sequence that overlapped with the *MEGF1* cDNA with 100% homology and extended the 5' end of the gene. The third RACE product only generated 130bp of sequence despite the PCR product being sized at ~240bp. This suggested the true 5' end of the *MEGF1* gene had been determined. In total, 1256bp of 5' RACE sequence had been generated and added to the 5' end of the *MEGF1* gene.

Figure 7.7a

Representative DHPLC analysis of exon 23e of the *PDGFRß* gene. A heterozygous pattern was observed in patient 11 (a) at 66°C. The amplified PCR product from the patient DNA was mixed with the amplified PCR product from the wild-type DNA. A large arrow indicates the homoduplex peak. A small arrow indicates the heteroduplex peak.

a)



A single, homozygous peak was observed in the normal control (b) at 66°C.

b)

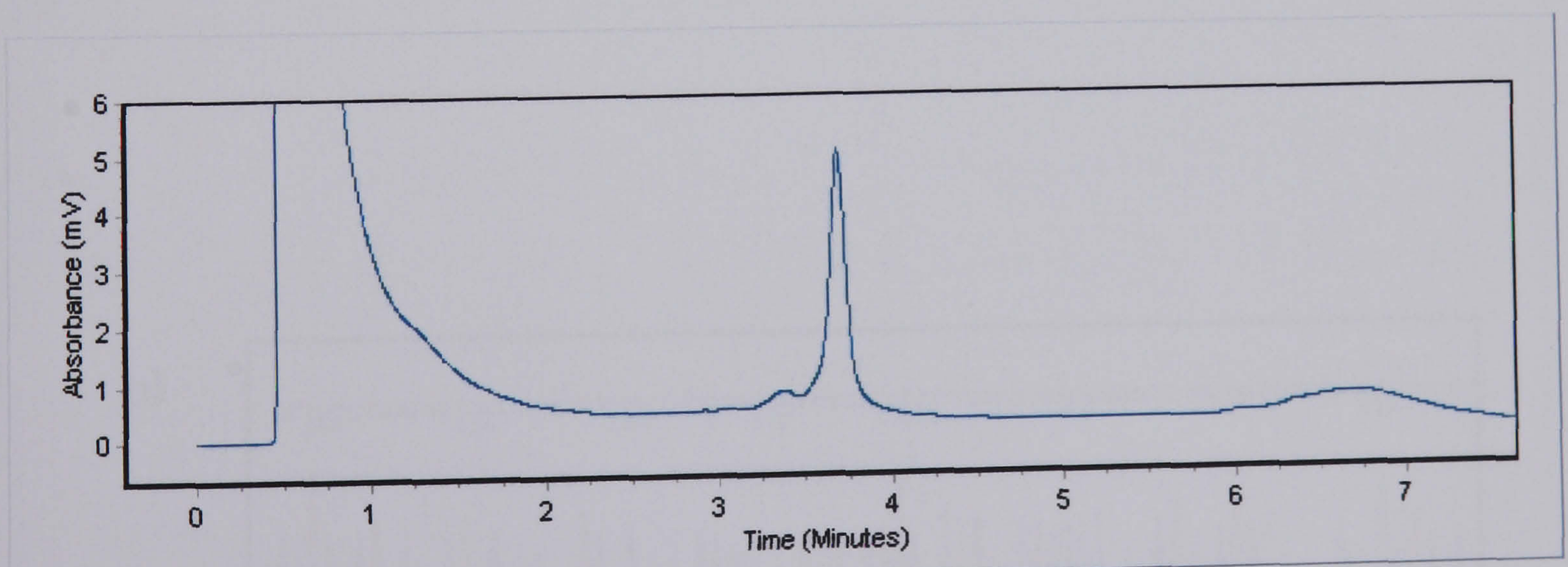
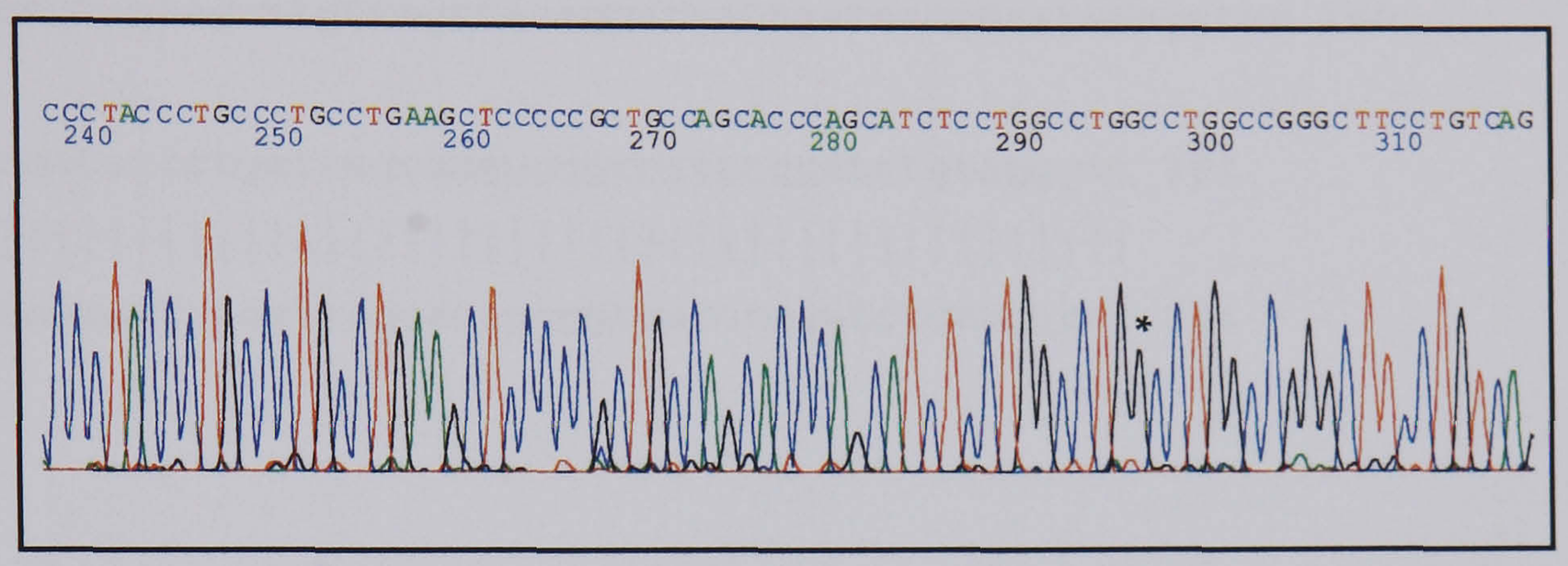


Figure 7.7b

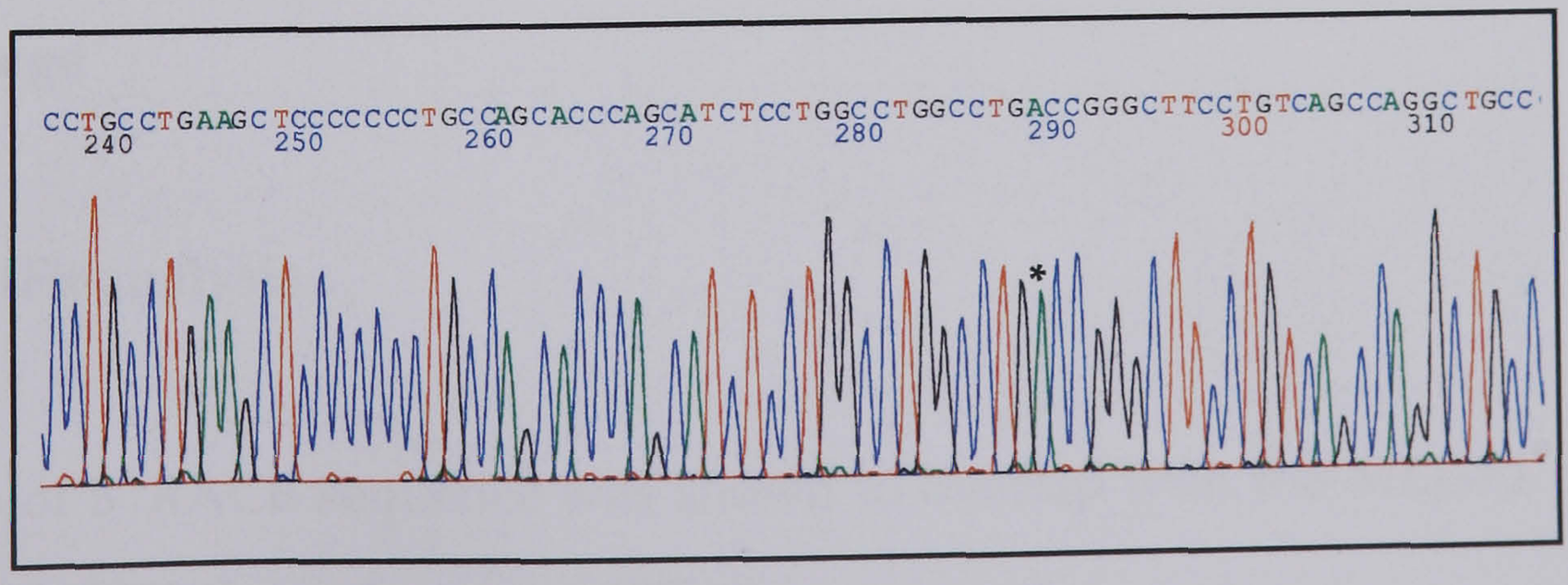
Representative sequence analysis of exon 23e of the *PDGFRß* gene. A G was observed in patient 11 (a) at nucleotide 72432. This was consistent with the published sequence.

a)



An A was observed in the normal control (b) at nucleotide 72432, suggesting patient 11 and the normal control were homozygous for alternate alleles in exon 23e of the *PDGFRß* gene. An asterisk (*) indicates the position of the alternate alleles.

b)



```
  1 cttaaatctgtctcagtaatggtatcacctagaccaccaacatatagtgt 50
    ||||||||||||||||||||||||||||||||||||||||||||||||||
  1 cttaaatctgtctcagtaatggtatcacctagaccaccaacatatagtgt 50
            .         .         .         .         .
 51 ggtgatagttttatcctctggtgggtccagccgaggcattgttgaagccc 100
    ||||||||||||||||||||||||||||||||||||||||||||||||||
 51 ggtgatagttttatcctctggtgggtccagccgaggcattgttgaagccc 100
            .         .         .         .         .
101 gctttagaagcttgtcagctacaggatcattgattccgtaataacggtct 150
    ||||||||||||||||||||||||||||||||||||||||||||||||||
101 gctttagaagcttgtcagctacaggatcattgattccgtaataacggtct 150
            .         .         .         .
151 ttaatattctgatcagcaaggggtcatctggatctgtaggctt 194
    |||||||||||||||||||||||||||||||||||||||||||
151 ttaatattctgatcagcaaggggtcatctggatctgtaggctt 194
```

Figure 7.8

Representative BestFit analysis of patient 13 (top) and the published sequence (bottom) from exon 8 of the ENSG00000086589 novel gene. The analysis shows no ambiguities between the two sequences suggesting no mutations or polymorphisms exist in this patient.

7.3.10 Database analysis using the genetic Computer Group (GCG) software package

7.3.10.1 BestFit analysis

The 1256bp of 5' RACE sequence was shown to overlap with the *MEGF1* cDNA sequence with 100% homology over 135bp.

7.3.10.2 BLAST analysis

A BlastX protein homology searches utilising the Mammalian sequences database showed the 1256bp RACE sequence to have 100% homology with the *Homo sapiens* protocadherin *FAT2* mRNA (human homologue of the *Drosophila* tumour suppressor gene, *fat2*) – GenBank accession number AF231022, over a 135bp overlap (*MEGF1* cDNA overlap). A BlastX protein homology search utilising the Genome sequences database showed the 1256bp RACE sequence to have 99% homology with the AC011374 *Homo sapiens* chromosome 5 clone CTB-113P19, working draft sequence, 37 unordered pieces over the whole 1256bp, see Figure 7.9. The Ensembl program had previously shown the *MEGF1* gene to map within contig AC011374 at 5q32.

7.3.10.3 Translate

The 1256bp RACE sequence could not be translated into one of the six open reading frames, suggesting the sequence was the 5' UTR of the *MEGF1* gene.

7.3.11 Genomic PCR

The GenBank database has shown the genomic sequence of contig AC011374 to be in 37 unordered pieces. Therefore, a genomic PCR was carried out to determine if the sequence of contig AC011374, upstream of the 5' end of the *MEGF1* gene, was in the correct order. Two overlapping PCR products (580bp and 547bp) were generated when genomic DNA was amplified across the genomic sequence of contig AC011374 containing the *MEGF1* gene, from the 5' RACE sequence. Direct sequencing of the two PCR products generated 580bp and 547bp of sequence respectively that showed 100% homology with the genomic sequence of contig AC011374 over the whole 1127bp. This confirmed the sequence upstream of the 5' end of the *MEGF1* gene was correct according to the GenBank database.

Figure 7.9 **BlastX analysis of 5' RACE sequence (top) and the *Homo sapiens* chromosome 5 working draft sequence, contig AC011374 (bottom)**

```
RACE:   1      taagtcatttacctttattatttcactgaattttcacaacaatcctaagagattgatgat 60
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
AC0113:70985   taagtcatttacctttattatttcactgaattttcacaacaatcctaagagattgatgat 71044

RACE:   61     tttcttatccttattctacagaaggtgaaatgaaggctcatgaggttaaataacttgctc 120
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
AC0113:71045   tttcttatccttattctacagaaggtgaaatgaaggctcatgaggttaaataacttgctc 71104

RACE:   121    taaaagtcacctacctagcaaatgacagaggcagggctgacagccaggcagcctgatgcc 180
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
AC0113:71105   taaaagtcacctacctagcaaatgacagaggcagggctgacagccaggcagcctgatgcc 71164

RACE:   181    acgcctgggatcttggtcactttgctttcctgcagcctctgcatgtgcggaaacagctca 240
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
AC0113:71165   acgcctgggatcttggtcactttgctttcctgcagcctctgcatgtgcggaaacagctca 71224

RACE:   241    gtttggctagagctcagccccagtggaaaggccttgtaggaaggagagagaatagtagtt 300
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
AC0113:71225   gtttggctagagctcagccccagtggaaaggccttgtaggaaggagagagaatagtagtt 71284

RACE:   301    tagaaccagatagctaaacttggttcctggctaagagttgggactatctcaccatcgccc 360
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
AC0113:71285   tagaaccagatagctaaacttggttcctggctaagagttgggactatctcaccatcgccc 71344

RACE:   361    aaggagtcacagaaagtttttgggaaaaaaagagtaaagagttcaggactgtgttctagct 420
               |||||||||||||||||||||||||||||||||| ||||||||||||||||||||||||
AC0113:71345   aaggagtcacagaaagtttttgggaaaaaaagagtaa-gagttcaggactgtgttctagct 71403

RACE:   421    gctgagtgtctttgagagagcatttttgaactagtctgctattcttcagccacggggttt 480
               |||||||||||||||||||||||||| ||||||||||||||||||||||||||||||||
AC0113:71404   gctgagtgtctttgagagagcatttt-gaactagtctgctattcttcagccacggggttt 71462

RACE:   481    ctcaaccatggcactattgacattggggctggataattctgttgcagggctgatatggtt 540
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
AC0113:71463   ctcaaccatggcactattgacattggggctggataattctgttgcagggctgatatggtt 71522

RACE:   541    tggctgtgttcccacccaaatctcatcttgaattgtggctcccataatcccccagtatca 600
               |||||||||||||||||||||||||||||||||||||||||||||||||||||| ||||||
AC0113:71523   tggctgtgttcccacccaaatctcatcttgaattgtggctcccataatcccccctgtatca 71582

RACE:   601    tgggagggatccagtgggaggtaattgaatcataggggtgtgttttcccatgctgttct 660
               |||||||||||||||||||||||||||||| |||||||||||||| ||||||||||||
AC0113:71583   tgggagggatccagtgggaggtaattgaatcatggggtgtgttttctcatgctgttct 71642

RACE:   661    cgtgatagtgaataagtctcatgagatctgatggttttacaaaggggagttcccctgcct 720
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
AC0113:71643   cgtgatagtgaataagtctcatgagatctgatggttttacaaaggggagttcccctgcct 71702

RACE:   721    acgcgctcttgcctgctgccaagtaagatatgactttgcttttcctttgccttctgccat 780
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
AC0113:71703   acgcgctcttgcctgctgccaagtaagatatgactttgcttttcctttgccttctgccat 71762

RACE:   781    gattatggggcctccccagtcatgtggaacggtgagtccattaaacctcattcctttata 840
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
AC0113:71763   gattatggggcctccccagtcatgtggaacggtgagtccattaaacctcattcctttata 71822

RACE:   841    aattacccagtctcaggtatgtctttattctcagtgtgagaactgactcacacaaaggca 900
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
AC0113:71823   aattacccagtctcaggtatgtctttattctcagtgtgagaactgactcacacaaaggca 71882
```

```
RACE:  901    ctgtcctatgcattgtaggttgtttagcagtatccccagcctctacctgccagatgccat 960
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
AC0113:71883  ctgtcctatgcattgtaggttgtttagcagtatccccagcctctacctgccagatgccat 71942

RACE:  961    tagcaccaccacccccaagctgggtcagccaaaagtatcttcagacattgccaagtgtcc 1020
              |||||||||||| ||||||||||||||||||||||||||||||||||||||||||||||
AC0113:71943  tagcaccaccactcccaagctgggtcagccaaaagtatcttcagacattgccaagtgtcc 72002

RACE:  1021   cttgagggtcaaaatcacctccagttaagaatcacagcagtagaaaccatttttccccc 1080
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
AC0113:72003  cttgagggtcaaaatcacctccagttaagaatcacagcagtagaaaccatttttccccc 72062

RACE:  1081   taacttgcgccctgtcttttattttctgcccagggtttcgggagttttccaccatgacta 1140
              ||||||||| ||||||||||||||||||||||||||||||||||||||||||||||||||
AC0113:72063  taacttgcaccctgtcttttattttctgcccagggtttcgggagttttccaccatgacta 72122

RACE:  1141   ttgccctgctgggttttgccatattcttgctccattgtgcgacctgtgagaagcctctag 1200
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
AC0113:72123  ttgccctgctgggttttgccatattcttgctccattgtgcgacctgtgagaagcctctag 72182

RACE:  1201   aagggattctctcctcctctgcttggcacttcacacactgcccattacaatgccac 1256
              |||||||||||||||||||||||||||||||||||||| ||||||||||||||||
AC0113:72183  aagggattctctcctcctctgcttggcacttcacacact-cccattacaatgccac 72237
```

The 1256bp 5' RACE sequence has a 99% nucleotide match with the *Homo sapiens* chromosome 5 working draft sequence, contig AC011374.

7.4 Discussion

The completion of the draft sequence of chromosome 5 by the Human Genome Project has enabled a collaboration with the Sanger Centre, using the Ensembl program, to predict the number of genes mapping to the approximate 1.5Mb critical region of the 5q- syndrome at 5q31.3-q32. The Ensembl program has predicted, in total, thirty-six genes of which twenty-three are known and thirteen are predicted (novel). The *GSHPx-3*, *PDGFRß*, *MEGF1* and novel genes 145872 and 86589 represent candidates for the 5q- syndrome gene, and were analysed for mutations by DHPLC.

The *MEGF1* gene is the human homologue of the *Drosophila* tumour suppressor gene *fat2* (Nakayama *et al.*, 1998). The *fat2* gene in *Drosophila* encodes a novel member of the cadherin superfamily (Mahoney *et al.*, 1991). The cadherins function as calcium-dependent adhesion molecules. The *fat2* tumour suppressor gene was identified because recessive mutations in the *fat2* locus cause hyperplastic, tumour-like overgrowth of larval imaginal discs, defects in morphogenesis and differentiation and death during the pupal stage. Other members of the cadherin family have been shown to function as tumour suppressor genes in human cancer. An example is the *E-cadherin* tumour suppressor gene that is frequently inactivated by mutation in human breast and gastric cancer (Berx *et al.*, 1998)

Human homologues of other tumour suppressor genes from *Drosophila* have been shown to possess tumour suppressor activity. *STIM1* (where STIM is stromal interaction molecule) is a candidate tumour suppressor gene that maps to human chromosome 11p15.5, a region implicated in a variety of cancers, particularly embryonal rhabdomyosarcoma (Williams *et al.*, 2001).

To investigate the proposal that *MEGF1* may be associated with the development of the 5q- syndrome, we analysed ten patients with the 5q- syndrome and seven patients with AML, for mutations in the twenty-three coding exons of the *MEGF1* gene. No mutations were found in the twenty-three coding exons of the *MEGF1* gene in the seventeen patients with the 5q- syndrome/AML included in the study. Sixteen previously unidentified polymorphisms were identified in the patient and normal control DNA from the *MEGF1* exons analysed by DHPLC.

During this study, the *MEGF1* gene was shown to be inactivated by downregulation of gene expression in a number of patients with the 5q- syndrome and AML. Hypermethylation of the promoter region of tumour suppressor genes may lead to tumour suppressor gene inactivation in cancer. For example, Mancini *et al.*, (1999) established a methylation map of the promoter region of the *NF1* tumour suppressor gene, and demonstrated functional sensitivity for methylation at specific sites for the SP1 and CRE binding (CREB) proteins in the *NF1* regulatory region.

To identify the promoter and methylation status of the CpG islands within the promoter region of the *MEGF1* gene, we determined the true 5' end of the *MEGF1* gene and identified the sequence upstream of the 5' end. To date, the promoter region and CpG islands cannot be identified using database promoter programs. Experimental studies will need to be carried out in order to identify the *MEGF1* promoter and to establish the methylation map of the promoter region of the *MEGF1* gene.

The *GSHPx-3* gene, like the *HAH1* gene, has been thought to play a role in antioxidant defence in cancer. An example of an antioxidant that may play a role in tumourigenesis is the superoxide dismutase (*SOD2*) gene, located on

chromosome 6q. A study by Bravard *et al.*, (1998) showed *SOD2* to have a lower activity in human melanoma cell lines with deletions of the 6q arm compared to the same cell lines without the 6q deletion.

Ten patients with the 5q- syndrome were sequenced for mutations in the five coding exons of the *GSHPx-3* gene. No mutations were found in the five coding exons of the *GSHPx-3* gene in the ten patients with the 5q- syndrome included in the study. A previously unidentified polymorphism in exon 5 was seen in patient 3 and in normal control DNA.

The *PDGFRß* gene has been shown to have a proven role in leukaemia. *PDGFRß* is a receptor tyrosine kinase that is disrupted by the t(5:7), t(5:12), and t(5:14) in myeloid disorders, resulting in the fusion of *PDGFRß* to *HIP1*, *TEL/ETV6*, and *CEV14*, respectively (Kulkarni *et al.*, 2000). The identification of these fusion genes involving *PDGFRß* strengthens the association between myeloproliferative disorders and deregulated tyrosine kinases. Other members of the type III receptor tyrosine kinase family include *FMS* (colony-stimulating factor 1R) and stem cell tyrosine kinase 1 (*STK-1*). Normal expression of *STK-1* is limited to $CD34^+$ stem/progenitor cells (Carow *et al.*, 1996). However, in a study of primary bone marrow (BM) samples from patients with leukaemia, *STK-1* was found to be expressed at a higher level in human leukaemias including AML, T-ALL, B-lineage acute leukaemia, and blast crisis CML, than in normal BM controls. Moreover, the STK-1 protein was found to be overexpressed in the leukaemic BM samples, suggesting *STK-1* may play a role in the survival and/or proliferation of malignant clones in acute myeloid and lymphoid leukaemias (Carow *et al.*, 1996). The *PDGFRß* gene was analysed for mutations by DHPLC in ten patients with the 5q- syndrome and two patients with AML. No mutations were found in the twenty-three coding exons of the *PDGFRß* gene in the twelve patients included in

the study. Sixteen previously unidentified polymorphisms were identified in patient and normal control DNA.

The novel gene ENSG00000145872 was identified as a human mitochondrial homologue of the bacterial co-chaperone GrpE (Ensembl gene report, 2001). Mitochondria contain a set of molecular chaperones, including hsp70, which are essential for the import of proteins from the cytoplasm into the mitochondrial matrix (Hartl *et al.*, 1992). Novel gene 113696 was selected as a candidate for the 5q- syndrome gene as it was expressed in $CD34^+$ cells, and some proteins of known tumour suppressor genes have previously been seen to act as chaperones. The *p53* and *RB1* tumour suppressor genes, the most commonly inactivated genes in human cancer, have been shown to act as powerful negative regulators of cell division (Lane *et al.*, 1993). The *RB1* gene achieves this by complexing to a variety of specific transcription factors and then inactivating their function. The capacity of the RB1 protein to bind these factors is regulated by phosphorylation. The RB1 proteins can therefore be seen to act as a chaperone for these factors (Lane *et al.*, 1993). The p53 protein has also been shown to regulate transcription, but may also be regulated by its interaction with members of the hsp70 chaperone family (Lane *et al.*, 1993).

To investigate the proposal that novel gene ENSG00000145872 may be mutated in the 5q- syndrome, we analysed ten patients with the 5q- syndrome and two patients with AML, for mutations in the three coding exons of 113696. No mutations were found in the three coding exons of the novel gene in the twelve patients included in the study.

Novel gene ENSG00000086589 has been shown to have the RNA-binding domain (RNA recognition motif), RNP-1 (Ensembl gene report, 2001). This novel gene

was selected for mutation analysis as it is expressed in $CD34^+$ cells and there have been reports of tumour suppressor genes encoding proteins that contain these binding domains. The *WT1* tumour suppressor gene encodes four C2H2 zinc finger-containing proteins critical for normal mammalian urogenital development (Kennedy *et al.*, 1996). *WT1* can bind specific DNA targets within the promoters of many genes and both transcriptional repression and activation domains have been identified (Kennedy *et al.*, 1996). Therefore, it has been assumed that regulation of transcription is the basis of *WT1* tumour suppressor activity.

We therefore decided to analyse novel gene ENSG00000086589 by DHPLC in ten patients with the 5q- syndrome and two patients with AML, for mutations in the twelve coding exons of 86589. No mutations were found in the twelve patients included in the study. Five previously unidentified polymorphisms were identified in patient and control DNA.

Results from this study have shown DHPLC using the WAVE™ DNA Fragment Analysis System to be an accurate mutation detection technique. The major advantages it has over other screening methods is its sensitivity, which we found to be 100%, a major reduction in laboratory time, and a reduced number of samples to be sequenced. One disadvantage is the necessity for high-fidelity PCR, although that is true for all mutation detection techniques. Its major disadvantage is the cost to the researcher in the maintenance and running of the machine which is more expensive than previous mutation detection techniques.

Chapter 8

Conclusion

The 5q- syndrome is a myelodysplastic disorder characterised by refractory anaemia, hypolobulated micromegakaryocytic hyperplasia and a clonal cytogenetic anomaly consisting of an interstitial deletion of the long arm of chromosome 5 (5q-) (Mathew *et al.*, 1993). It is widely believed that a gene(s) located on 5q may function as a leukaemia suppressor gene (Le Beau, 1992).

In order to identify the putative 5q- syndrome tumour suppressor gene, we used the EST resource to generate a transcription map of the approximate 5Mb critical region of gene loss at 5q31-q33, flanked by the genes *FGF1* and *IL12ß*. In the first instance we identified, isolated and mapped ten novel coding sequences to the YAC contig spanning the critical region. This included the cloning of novel gene, *C5orf4*, and the identification and mapping of the human synaptopodin and dynactin *p62* genes, see Figure 8.1. This was followed by the identification and localisation of the human homologues of the *Drosophila melanogaster RMSA-1*, *Saccharomyces cerevisiae CDC60*, and Goliath protein genes, and the localisation of known human genes *PP2A*, *tpr1*, *PPP1R2*, and *HAH1* to the transcript map, see Figure 8.1. These known and novel genes have contributed to the overall mapping of this genomic region and represent candidates for the 5q- syndrome gene. During the course of this study, however, several of these genes were eliminated from further analysis when the commonly deleted region of the 5q- syndrome was narrowed to approximately 1.5Mb at 5q31.3-q32, flanked by the DNA marker D5S413 and the *GLRA1* gene.

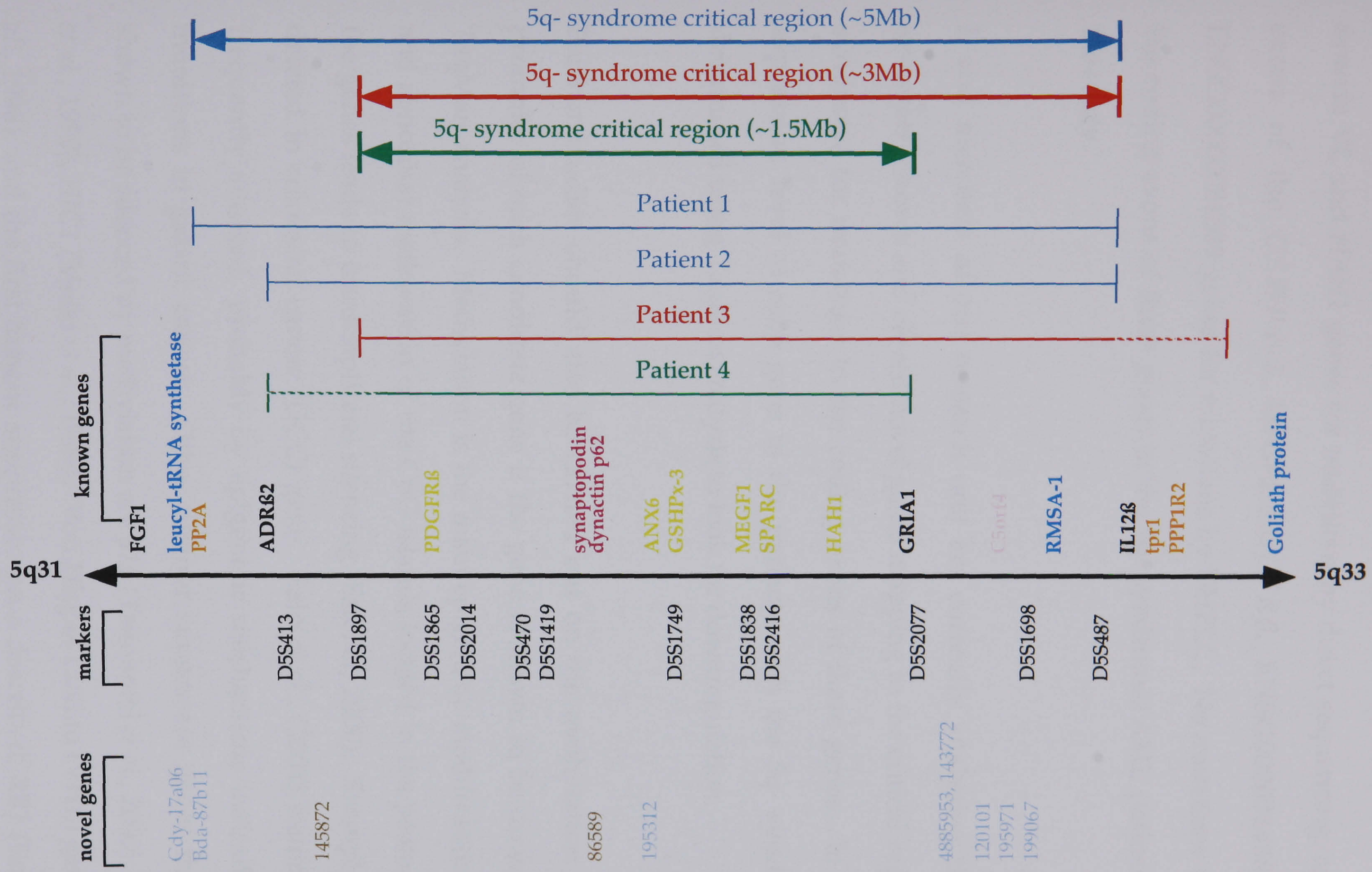


Figure 8.1 Transcription map of the critical region (CR) of the 5q- syndrome. The map shows the patients that defined the 5MB CR (blue), the 3Mb CR (red), and the 1.5Mb CR (green). The genes that flank the CR breakpoints (black), novel gene C5orf4 (pink) cloned in this study, novel genes (lilac) identified in this study, known genes identified from novel ESTs (magenta), human homologues (turquoise) identified in this study, human genes (orange) localised to the CR, known genes selected for mutation studies (yellow), and novel genes selected for mutation studies (brown) are shown.

From the 1.5Mb critical region, we selected six known and two novel candidate genes for further analysis. We analysed the coding region/exons of the *SPARC*, *annexin* VI, and *HAH1* genes for mutations by direct sequencing, and the coding exons of the *GSHPx-3*, *MEGF1*, *PDGFRß*, ENSG00000145872, and the ENSG00000086589 genes for mutations by DHPLC. No mutations were found in the coding exons of these genes in the 5q- syndrome/AML patients included in the study.

I, and members of our research unit are currently analysing the remaining seventeen known and eleven novel genes mapping to the critical region of the 5q- syndrome for mutations in the coding exons of these genes. In addition, the expression level of each gene in the patients with the 5q- syndrome will be determined using TaqMan analysis to look for downregulation.

Further studies should also be carried out on the methylation status of the promoter of each candidate gene if the gene is shown to be downregulated by TaqMan analysis. Methylation is the main epigenetic modification in mammals and abnormal methylation of the CpG islands located in the promoter region of the genes leads to transcriptional silencing (Esteller, 2000). Examples include the deleted in colorectal cancer (*DCC*) gene. Sato *et al.*, (2001) found that *DCC* is frequently silenced, probably by epigenetic mechanisms instead of sequence mutations in gastric cancer. Other tumour suppressor genes that have been shown to be silenced by methylation are *p16* (Tannapfel *et al.*, 2000), *NF1* (Mancini *et al.*, 1999), *HIC1* (Melki *et al.*, 1999)/ Von Hippel-Lindau (VHL) gene (Clifford et *al.*, 1998), and the first tumour suppressor gene described, *RB1* (Robertson *et al.*, 2000).

During this study we found the *MEGF1* gene to be downregulated in a number of patients with the 5q- syndrome and AML compared to normal controls. The

identification of the *MEGF1* promoter to evaluate the methylation status of CpG islands within the promoter region is in progress.

It is possible that Knudson's two-hit hypothesis may not be relevant in the development of the 5q- syndrome and that haploinsufficiency may be the underlying mechanism. It is generally assumed that most mammalian genes are transcribed from both alleles. Hence, the diploid state of the genome offers the advantage that a loss-of-function mutation in one allele can be compensated for by the remaining wild-type allele of the same gene (Nutt and Busslinger, 1999). It is well known that the vast majority of human disease-causing genes are recessive, indicating that recessiveness is the 'default' state. However, a minority of genes are semi-dominant, as heterozygous loss-of-function mutation in these genes leads to phenotypic abnormalities. This condition is known as haploinsufficiency. Haploinsufficiency is believed to be the underlying mechanism in many diseases.

Song *et al.*, (1999) demonstrated that haploinsufficiency of the *AML1* gene is the genetic basis of a form of familial thrombocytopenia which predisposes the affected individuals to the development of acute myeloid leukaemia. p27Kip is a candidate human tumour suppressor protein as it is able to inhibit cyclin-dependent kinases and block cell proliferation (Fero *et al.*, 1998). However, a causal link between *p27* and tumour suppression has not been established as homozygous inactivating mutations of the *p27* gene in human tumours is a rare occurrence. Thus, *p27Kip1* does not fulfil Knudson's 'two-mutation' criterion for a tumour suppressor gene. Fero *et al.*, demonstrated that molecular analyses of tumours in *p27* heterozygous mice showed the remaining wild-type allele to be neither mutated nor silenced, suggesting *p27* is haploinsufficient for tumour suppression.

Knudson's hypothesis (the inactivation of two alleles) or haploinsufficiency are the two possibilities as the underlying mechanism in the 5q- syndrome. Extensive

studies need to be carried out on all candidate genes mapping to the approximate 1.5Mb critical region at 5q31.3-q32.

In recent years, the use of mouse models has greatly contributed to the understanding of the role of tumour suppressor gene function. Novel insights into the role of tumour suppressors in development, differentiation, cell cycle control, and tumour suppression have been obtained from the studies on these 'knockout' mice. In addition, such mice may serve as disease models for humans with inherited cancer predisposition syndromes. The advantage of many mouse tumour suppressor models is that they facilitate the study of the roles of tumour suppressor gene loss in tumour initiation and progression *in vivo*. Moreover, the extraction of primary cells from tumour suppressor-deficient mice has provided an important resource for *in vitro* studies on the role of targeted genes in cell cycle regulation, DNA damage response, regulation of apoptotic pathways, and preservation of genomic stability (Ghebranious and Donehower, 1998). For example, a knockout mouse has contributed to the understanding of the role of *p53* in tumour suppression. Mice homozygous for a deletion in the *p53* gene develop tumours at high frequency, providing essential evidence for the importance of *p53* as a tumour suppressor in several human cancers (Attardi and Jacks, 1999).

The ability to manipulate the mouse genome via overexpression, underexpression or deletion of genes using transgenic expression systems and embryonic stem cell (ES) technology has led to the identification and definition of the precise function of several tumour suppressor genes *in vivo*. This group includes mice with mutations in the *RB1* gene. In contrast to the role of the *RB1* gene in human retinoblastomas, mice heterozygous for a mutant RB allele do not develop retinoblastoma, but develop pituitary tumours instead (Kumar *et al.*, 1995). The tumour susceptibility phenotype of mutant mice has unveiled the tumour suppressor activity of specific genes that were not expected to have such a

function. Transgenic and knockout mice will have an increasingly important role in the identification of novel tumour suppressor genes (Kumar *et al.*, 1995). This could be important in the identification of the putative 5q- syndrome tumour suppressor gene. The generation of a knockout mice for the 5q- syndrome is currently in progress. This method will address either a 'one-hit' (haplolinsufficiency) or 'two-hit' hypothesis as the underlying mechanism in the pathogenesis of the 5q- syndrome.

The targeting of genes involved in the pathogenesis of cancer and disease is now being carried out on a global scale using DNA microarrays. The advent of cDNA microarray technology now allows the efficient measurement of expression for almost every gene in the human genome. Novel molecular-based sub-classes of tumours in breast carcinoma, colon carcinoma, lymphoma, leukaemia, and melanoma have been revealed using global expression profiling (Alizadeh *et al.*, 2001). DNA microarray analysis has already been shown to be of value in MDS. Despite the relatively high incidence of MDS in the elderly, differentiation of MDS from *de novo* AML still remains problematic. Through the use of oligonucleotide arrays, the gene encoding the protein Delta-like (*Dlk*) that is distantly related to the Delta-Notch family of signalling proteins, was found to be selectively expressed in patients with MDS compared to patients with AML and CML. Thus, *Dlk* could be the first candidate molecule to differentiate MDS from AML (Miyazato *et al.*, 2001). We are currently using DNA microarray technology to identify genes that may be over or underexpressed in patients with the 5q- syndrome.

The Human Genome Project originally was planned to last fifteen years, but effective resources and technological advances have accelerated the expected completion date to 2003. Several types of genome maps have already been completed, and a working draft of the entire human genome sequence was announced in June 2000, with analyses published in February 2001. It is most

probable that the availability of the complete annotated genomic sequence from human chromosome 5q will be the key in identifying and characterising the 5q- syndrome gene.

The identification of the 5q- syndrome gene will enable the study of its protein at a functional level. Proteomics has contributed greatly to the understanding of gene function in the post-genomic era. Proteomics can be divided into three main areas: (1) protein micro-characterisation for large-scale identification of proteins and their post translational modifications; (2) 'differential display' proteomics for comparison of protein levels with potential application in a wide range of diseases; and (3) studies of protein-protein interaction using techniques such as the yeast-two-hybrid system. Investigators have used the two-hybrid system to directly assay interactions between known proteins and to isolate novel interacting partners for a protein of interest. The identification of mutations in each partner of an interacting pair of proteins, which disrupt the interaction, can be useful for generating genetic tools for characterising *in vivo* function. The functional characterisation of some tumour suppressor genes have been ascertained using the yeast two-hybrid system, including the familial breast and ovarian cancer susceptibility genes, *BRAC1* and *BRAC2* (Sharan *et al.*, 1998). This system identified several murine *Brca1* and *Brca2* interacting proteins, including BARD1. Recently, mutations suggesting a role as a tumour suppressor have been identified in the *BARD1* gene in primary human tumours. The identification of molecules that interact with murine *Brca1* and *Brca2* has greatly enhanced our knowledge of how *BRCA1* and *BRCA2* may function as tumour suppressor genes.

Transfection studies using transformed NIH3T3 cells should also be carried out once the 5q- syndrome gene has been identified to observe any phenotypic changes the gene may cause. These studies have been carried out on other genes implicated in leukaemogenesis. For example, Kurokawa *et al.*, (1996) demonstrated that the AMLb splice variant of the *AML1* gene causes neoplastic

transformation of NIH3T3 cells. The elucidation of function of the 5q- syndrome gene could also be determined with the use of leukaemic cell lines.

In conclusion, I have been involved in the generation of a transcript map of the 5q- syndrome critical region. These newly assigned genes have contributed to the detailed mapping of the region and have been investigated as candidate genes. Once the 5q- syndrome gene has been identified, it should be investigated to ascertain its implications, if any, in the pathogenesis of a wide spectrum of human cancers and leukaemias, as for the *p53* tumour suppressor gene.

Publications

Some of the data in this thesis has been published;

Boultwood, J., C. Fidler, **A. J. Strickson**, F. Watkins, M. Kostrzewa, R. J. Jaju, U. Muller and J. S. Wainscoat (2000). Transcription mapping of the 5q- syndrome critical region: cloning of two novel genes and sequencing, expression, and mapping of a further six novel cDNAs. Genomics **66**: 26-34.

Boultwood, J., **A. J. Strickson**, E. W. Jabs, J. F. Cheng, C. Fidler and J. S. Wainscoat (2000). Physical mapping of the human ATX1 homologue (HAH1) to the critical region of the 5q- syndrome within 5q32, and immediately adjacent to the SPARC gene. Hum Genet **106**: 127-9.

Boultwood, J., C. Fidler, P. Soularue, **A. J. Strickson**, M. Kostrzewa, R. J. Jaju, F. E. Cotter, N. Fairweather, A. P. Monaco, U. Muller, M. Lovett, E. W. Jabs, C. Auffray and J. S. Wainscoat (1997). Novel genes mapping to the critical region of the 5q- syndrome. Genomics **45**: 88-96.

Fidler, C., **A. J. Strickson**, J. Boultwood and J. S. Wainscoat (2000). Mutation analysis of the SPARC gene in the 5q-syndrome. Am J Haematol **64**: 324

Strickson, A. J., and C. Fidler (2002). Mutation analysis of cancer using automated sequencing. Molecular analysis of Cancer. Totowa, New Jersey. Humana Press. 171-177.

References

Adams, M. D., J. M. Kelley, J. D. Gocayne, M. Dubnick, M. H. Polymeropoulos, H. Xiao, C. R. Merril, A. Wu, B. Olde, R. F. Moreno and et al. (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. Science **252**: 1651-6.

Adamson, D. J., A. A. Dawson, B. Bennett, D. J. King and N. E. Haites (1995). p53 mutation in the myelodysplastic syndromes. Br J Haematol **89**: 61-6.

Alizadeh, A. A., D. T. Ross, C. M. Perou and M. van de Rijn (2001). Towards a novel classification of human malignancies based on gene expression patterns. J Pathol **195**: 41-52.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman (1990). Basic local alignment search tool. J Mol Biol **215**: 403-10.

Attardi, L. D. and T. Jacks (1999). The role of p53 in tumour suppression: lessons from mouse models. Cell Mol Life Sci **55**: 48-63.

Auffray, C., G. Behar, F. Bois, C. Bouchier, C. Da Silva, M. D. Devignes, S. Duprat, R. Houlgatte, M. N. Jumeau, B. Lamy and et al. (1995). [IMAGE: molecular integration of the analysis of the human genome and its expression]. C R Acad Sci III **318**: 263-72.

Azim, A. C., J. H. Knoll, S. M. Marfatia, D. J. Peel, P. J. Bryant and A. H. Chishti (1995). DLG1: chromosome location of the closest human homologue of the Drosophila discs large tumor suppressor gene. Genomics **30**: 613-6.

Baffa, R., R. Santoro, F. Bullrich, B. Mandes, H. Ishii and C. M. Croce (2000). Definition and refinement of chromosome 8p regions of loss of heterozygosity in gastric cancer. Clin Cancer Res **6**: 1372-7.

Banfi, S., G. Borsani, A. Bulfone and A. Ballabio (1997). Drosophila-related expressed sequences. Hum Mol Genet **6**: 1745-53.

Bardeesy, N. and J. Pelletier (1998). Overlapping RNA and DNA binding domains of the wt1 tumor suppressor gene product. Nucleic Acids Res **26**: 1784-92.

Baysal, B. E., E. M. van Schothorst, J. E. Farr, M. R. James, P. Devilee and C. W. Richard, 3rd (1997). A high-resolution STS, EST, and gene-based physical map of the hereditary paraganglioma region on chromosome 11q23. Genomics **44**: 214-21.

Bench, A. J., E. P. Nacheva, T. L. Hood, J. L. Holden, L. French, S. Swanton, K. M. Champion, J. Li, P. Whittaker, G. Stavrides, A. R. Hunt, B. J. Huntly, L. J. Campbell, D. R. Bentley, P. Deloukas and A. R. Green (2000). Chromosome 20 deletions in myeloid malignancies: reduction of the common deleted region, generation of a PAC/BAC contig and identification of candidate genes. UK Cancer Cytogenetics Group (UKCCG). Oncogene **19**: 3902-13.

Berx, G., K. F. Becker, H. Hofler, F. Roy (1998). Mutations of the human E-cadherin (CDH1) gene. Hum mutat **12**: 226-237.

Bezieau, S., M. C. Devilder, G. Rondeau, E. Cadoret, J. P. Moisan and I. Moreau (1998). Assignment of 48 ESTs to chromosome 13 band q14.3 and expression pattern for ESTs located in the core region deleted in B-CLL. Genomics **52**: 369-73.

Bharaj, B. S., K. Angelopoulou and E. P. Diamandis (1998). Rapid sequencing of the p53 gene with a new automated DNA sequencer. Clin Chem **44**: 1397-403.

Blanquet, V., C. Turleau, M. S. Gross, M. Goossens and C. Besmond (1993). Identification of germline mutations in the RB1 gene by denaturant gradient gel electrophoresis and polymerase chain reaction direct sequencing. Hum Mol Genet **2**: 975-9.

Blatch, G. L. and M. Lassle (1999). The tetratricopeptide repeat: a structural motif mediating protein- protein interactions. Bioessays **21**: 932-9.

Boguski, M. S. (1995). The turning point in genome research. Trends Biochem Sci **20**: 295-6.

Boguski, M. S. and G. D. Schuler (1995). ESTablishing a human transcript map. Nat Genet **10**: 369-71.

Bolufer, P., G. F. Sanz, E. Barragan, M. A. Sanz, J. Cervera, E. Lerma, L. Senent, I. Moreno and M. D. Planelles (2000). Rapid quantitative detection of BCR-ABL transcripts in chronic myeloid leukemia patients by real-time reverse transcriptase polymerase-chain reaction using fluorescently labeled probes. Haematologica **85**: 1248-54.

Borkhardt, A., S. Bojesen, O. A. Haas, U. Fuchs, D. Bartelheimer, I. F. Loncarevic, R. M. Bohle, J. Harbott, R. Repp, U. Jaeger, S. Viehmann, T. Henn, P. Korth, D. Scharr and F. Lampert (2000). The human GRAF gene is fused to MLL in a unique t(5;11)(q31;q23) and both alleles are disrupted in three cases of myelodysplastic syndrome/acute myeloid leukemia with a deletion 5q. Proc Natl Acad Sci U S A **97**: 9168-73.

Boultwood, J. and C. Fidler (1995). Chromosomal deletions in myelodysplasia. Leuk Lymphoma **17**: 71-8.

Boultwood, J., C. Fidler, S. Lewis, S. Kelly, H. Sheridan, T. J. Littlewood, V. J. Buckle and J. S. Wainscoat (1994). Molecular mapping of uncharacteristically small 5q deletions in two patients with the 5q- syndrome: delineation of the critical region on 5q and identification of a 5q- breakpoint. Genomics **19**: 425-32.

Boultwood, J., C. Fidler, S. Lewis, A. MacCarthy, H. Sheridan, S. Kelly, D. Oscier, V. J. Buckle and J. S. Wainscoat (1993). Allelic loss of IRF1 in myelodysplasia and acute myeloid leukemia: retention of IRF1 on the 5q- chromosome in some patients with the 5q- syndrome. Blood **82**: 2611-6.

Boultwood, J., C. Fidler, P. Soularue, A. J. Strickson, M. Kostrzewa, R. J. Jaju, F. E. Cotter, N. Fairweather, A. P. Monaco, U. Muller, M. Lovett, E. W. Jabs, C. Auffray and J. S. Wainscoat (1997). Novel genes mapping to the critical region of the 5q- syndrome. Genomics **45**: 88-96.

Boultwood, J., C. Fidler, A. J. Strickson, F. Watkins, M. Kostrzewa, R. J. Jaju, U. Muller and J. S. Wainscoat (2000). Transcription mapping of the 5q- syndrome critical region: cloning of two novel genes and sequencing, expression, and mapping of a further six novel cDNAs. Genomics **66**: 26-34.

Boultwood, J., A. J. Strickson, E. W. Jabs, J. F. Cheng, C. Fidler and J. S. Wainscoat (2000). Physical mapping of the human ATX1 homologue (HAH1) to the critical region of the 5q- syndrome within 5q32, and immediately adjacent to the SPARC gene. Hum Genet **106**: 127-9.

Boyum, A. (1984). Separation of lymphocytes, granulocytes, and monocytes from human blood using iodinated density gradient media. Methods Enzymol **108**: 88-102.

Bravard, A., C. Cherbonnel-Lasserre, M. Reillaudou, J. Beaumatin, B. Dutrillaux and C. Luccioni (1998). Modifications of the antioxidant enzymes in relation to chromosome imbalances in human melanoma cell lines. Melanoma Res **8**: 329-35.

Bronner, C. E., S. M. Baker, P. T. Morrison, G. Warren, L. G. Smith, M. K. Lescoe, M. Kane, C. Earabino, J. Lipford, A. Lindblom and et al. (1994). Mutation in the DNA mismatch repair gene homologue hMLH1 is associated with hereditary non-polyposis colon cancer. Nature **368**: 258-61.

Buckler, A. J., D. D. Chang, S. L. Graw, J. D. Brook, D. A. Haber, P. A. Sharp and D. E. Housman (1991). Exon amplification: a strategy to isolate mammalian genes based on RNA splicing. Proc Natl Acad Sci U S A **88**: 4005-9.

Burston, S. G. and A. R. Clarke (1995). Molecular chaperones: physical and mechanistic properties. Essays Biochem **29**: 125-36.

Carow, C. E., M. Levenstein, S. H. Kaufmann, J. Chen, S. Amin, P. Rockwell, L. Witte, M. J. Borowitz, C. I. Civin and D. Small (1996). Expression of the hematopoietic growth factor receptor FLT3 (STK- 1/Flk2) in human leukemias. Blood **87**: 1089-96.

Castro, P. D., J. C. Liang and L. Nagarajan (2000). Deletions of chromosome 5q13.3 and 17p loci cooperate in myeloid neoplasms. Blood **95**: 2138-43.

Cerutti, P. A. (1985). Prooxidant states and tumor promotion. Science **227**: 375-81.

Chadwick, R. B., G. L. Jiang, G. A. Bennington, B. Yuan, C. K. Johnson, M. W. Stevens, T. H. Niemann, P. Peltomaki, S. Huang and A. de la Chapelle (2000). Candidate tumor suppressor RIZ is frequently involved in colorectal carcinogenesis. Proc Natl Acad Sci U S A **97**: 2662-7.

Chenchik, A., L. Diachenko, F. Moqadam, V. Tarabykin, S. Lukyanov and P. D. Siebert (1996). Full-length cDNA cloning and determination of mRNA 5' and 3' ends by amplification of adaptor-ligated cDNA. Biotechniques **21**: 526-34.

Chu, F. F., R. S. Esworthy, J. H. Doroshow, K. Doan and X. F. Liu (1992). Expression of plasma glutathione peroxidase in human liver in addition to kidney, heart, lung, and breast in humans and rodents. Blood **79**: 3233-8.

Clark, D. M., S. E. Moss, N. A. Wright and M. J. Crumpton (1991). Expression of annexin VI (p68, 67 kDa-calelectrin) in normal human tissues: evidence for developmental regulation in B- and T-lymphocytes. Histochemistry **96**: 405-12.

Clifford, S. C., A. H. Prowse, N. A. Affara, C. H. Buys and E. R. Maher (1998). Inactivation of the von Hippel-Lindau (VHL) tumour suppressor gene and allelic losses at chromosome arm 3p in primary renal cell carcinoma: evidence for a VHL-independent pathway in clear cell renal tumourigenesis. Genes Chromosomes Cancer **22**: 200-9.

Creutz, C. E., S. Moss, J. M. Edwardson, I. Hide and B. Gomperts (1992). Differential recognition of secretory vesicles by annexins. European Molecular Biology Organization Course "Advanced Techniques for Studying Secretion". Biochem Biophys Res Commun **184**: 347-52.

Crompton, M. R., S. E. Moss and M. J. Crumpton (1988). Diversity in the lipocortin/calpactin family. Cell **55**: 1-3.

Cross, C. E., B. Halliwell, E. T. Borish, W. A. Pryor, B. N. Ames, R. L. Saul, J. M. McCord and D. Harman (1987). Oxygen radicals and human disease. Ann Intern Med **107**: 526-45.

Crowder, S., J. Holton and T. Alber (2001). Covariance analysis of RNA recognition motifs identifies functionally linked amino acids. J Mol Biol **310**: 793-800.

Crumpton, M. J. and J. R. Dedman (1990). Protein terminology tangle. Nature **345**: 212.

Den Dunnen, J. T. and G. J. Van Ommen (1999). The protein truncation test: A review. Hum Mutat **14**: 95-102.

Del Mastro, R. G. and M. Lovett (1997). In: Boultwood J, (ed.) Gene Isolation and Mapping Protocols. Totowa, New Jersey: Humana Press: 201-210.

Dewald, G. W., M. P. Davis, R. V. Pierre, J. R. O'Fallon and H. C. Hoagland (1985). Clinical characteristics and prognosis of 50 patients with a myeloproliferative syndrome and deletion of part of the long arm of chromosome 5. Blood **66**: 189-97.

Doyle, L. A., W. Yang, A. K. Rishi, Y. Gao and D. D. Ross (1996). H19 gene overexpression in atypical multidrug-resistant cells associated with expression of a 95-kilodalton membrane glycoprotein. Cancer Res **56**: 2904-7.

Dreher, D. and A. F. Junod (1996). Role of oxygen free radicals in cancer development. Eur J Cancer **32A**: 30-8.

Dunham, I., N. Shimizu, B. A. Roe, S. Chissoe, A. R. Hunt, J. E. Collins, R. Bruskiewich, D. M. Beare, M. Clamp, L. J. Smink, R. Ainscough, J. P. Almeida, A. Babbage, C. Bagguley, J. Bailey, K. Barlow, K. N. Bates, O. Beasley, C. P. Bird, S. Blakey, A. M. Bridgeman, D. Buck, J. Burgess, W. D. Burrill, K. P. O'Brien and et al. (1999). The DNA sequence of human chromosome 22. Nature **402**: 489-95.

Elgar, G., M. S. Clark, S. Meek, S. Smith, S. Warner, Y. J. Edwards, N. Bouchireb, A. Cottage, G. S. Yeo, Y. Umrania, G. Williams and S. Brenner (1999). Generation and analysis of 25 Mb of genomic DNA from the pufferfish Fugu rubripes by sequence scanning. Genome Res **9**: 960-71.

Esposito, T., F. Gianfrancesco, A. Ciccodicola, M. D'Esposito, R. Nagaraja, R. Mazzarella, M. D'Urso and A. Forabosco (1997). Escape from X inactivation of two new genes associated with DXS6974E and DXS7020E. Genomics **43**: 183-90.

Esteller, M. (2000). Epigenetic lesions causing genetic lesions in human cancer: promoter hypermethylation of DNA repair genes. Eur J Cancer **36**: 2294-300.

Fero, M. L., E. Randel, K. E. Gurley, J. M. Roberts and C. J. Kemp (1998). The murine gene p27Kip1 is haplo-insufficient for tumour suppression. Nature **396**: 177-80.

Fidler, C., M. Nakayama, E. W. Jabs, J. F. Cheng, A. J. Strickson, O. Ohara, J. Boultwood and J. S. Wainscoat (2001). Physical mapping of the MEGF1 gene, human homologue of the Drosophila tumour suppressor gene fat, to the critical region of the 5q- syndrome. GeneScreen **1**: 165-167

Fidler, C., J. S. Wainscoat and J. Boultwood (1999). The human POP2 gene: identification, sequencing, and mapping to the critical region of the 5q- syndrome. Genomics **56**: 134-6.

Fidler , C., and J. Boultwood (1997). Isolation of cDNAs using the YAC hybridisation screen method . In: Boultwood J, (ed.) Gene Isolation and Mapping Protocols. Totowa, New Jersey: Humana Press: 201-210.

Fodde, R. and M. Losekoot (1994). Mutation detection by denaturing gradient gel electrophoresis (DGGE). Hum Mutat **3**: 83-94.

Fodde, R., R. van der Luijt, J. Wijnen, C. Tops, H. van der Klift, I. van Leeuwen-Cornelisse, G. Griffioen, H. Vasen and P. M. Khan (1992). Eight novel inactivating germ line mutations at the APC gene identified by denaturing gradient gel electrophoresis. Genomics **13**: 1162-8.

Fourth International Workshop on Chromosomes in Leukaemia, 1984.

Ghebranious, N. and L. A. Donehower (1998). Mouse models in tumor suppression. Oncogene **17**: 3385-400.

Giarola, M., L. Stagi, S. Presciuttini, P. Mondini, M. T. Radice, P. Sala, M. A. Pierotti, L. Bertario and P. Radice (1999). Screening for mutations of the APC gene in 66 Italian familial adenomatous polyposis patients: evidence for phenotypic differences in cases with and without identified mutation. Hum Mutat **13**: 116-23.

Glavac, D. and M. Dean (1993). Optimization of the single-strand conformation polymorphism (SSCP) technique for detection of point mutations. Hum Mutat **2**: 404-14.

Golub, T. R., G. F. Barker, M. Lovett and D. G. Gilliland (1994). Fusion of PDGF receptor beta to a novel ets-like gene, tel, in chronic myelomonocytic leukemia with t(5;12) chromosomal translocation. Cell **77**: 307-16.

Goodrow, T. L. (1996). One decade of comparative molecular carcinogenesis. Prog Clin Biol Res **395**: 57-80.

Gordon, M. S. (1999). Advances in supportive care of myelodysplastic syndromes. Semin Hematol **36**: 21-4.

Graves, J. A. (1998). Background and Overview of Comparative Genomics. Ilar J **39**: 48-65.

Grimwade, D. J., J. Stephenson, C. De Silva, R. G. Dalton and G. J. Mufti (1993). Familial MDS with 5q- abnormality. Br J Haematol **84**: 536-8.

Groden, J., A. Thliveris, W. Samowitz, M. Carlson, L. Gelbert, H. Albertsen, G. Joslyn, J. Stevens, L. Spirio, M. Robertson and et al. (1991). Identification and characterization of the familial adenomatous polyposis coli gene. Cell **66**: 589-600.

Gronwald, R. G., F. J. Grant, B. A. Haldeman, C. E. Hart, P. J. O'Hara, F. S. Hagen, R. Ross, D. F. Bowen-Pope and M. J. Murray (1988). Cloning and expression of a cDNA coding for the human platelet-derived growth factor receptor: evidence for more than one receptor class. Proc Natl Acad Sci U S A **85**: 3435-9.

Gross, E., N. Arnold, J. Goette, U. Schwarz-Boeger and M. Kiechle (1999). A comparison of BRCA1 mutation analysis by direct sequencing, SSCP and DHPLC. Hum Genet **105**: 72-8.

Guerardel, C., S. Deltour, S. Pinte, D. Monte, A. Begue, A. K. Godwin and D. Leprince (2001). Identification in the human candidate tumor suppressor gene HIC-1 of a new major alternative TATA-less promoter positively regulated by p53. J Biol Chem **276**: 3078-89.

Halliwell, B. (1987). Oxidants and human disease: some new concepts. Faseb J **1**: 358-64.

Hanson, I. M. and J. Trowsdale (1991). Colinearity of novel genes in the class II regions of the MHC in mouse and human. Immunogenetics **34**: 5-11.

Hao, Y., T. Crenshaw, T. Moulton, E. Newcomb and B. Tycko (1993). Tumour-suppressor activity of H19 RNA. Nature **365**: 764-7.

Hartl, F. U., J. Martin and W. Neupert (1992). Protein folding in the cell: the role of molecular chaperones Hsp70 and Hsp60. Annu Rev Biophys Biomol Struct **21**: 293-322.

Hass, J., K. Mayer and H. D. Rott (2000). Tuberous sclerosis type 1: three novel mutations detected in exon 15 by a combination of HDA and TGGE. Hum Mutat **16**: 88.

Hayashi, K. and D. W. Yandell (1993). How sensitive is PCR-SSCP? Hum Mutat **2**: 338-46.

Heim, S., and F. Mitelman (1986). Chromosome abnormalities in the myelodysplastic syndromes. Clin Haematol **15**: 1003-1021.

Hohmann, S. and J. M. Thevelein (1992). The cell division cycle gene CDC60 encodes cytosolic leucyl-tRNA synthetase in Saccharomyces cerevisiae. Gene **120**: 43-9.

Horrigan, S. K., L. Bartoloni, M. C. Speer, N. Fulton, J. Kravarusic, R. Ramesar, J. M. Vance, L. H. Yamaoka and C. A. Westbrook (1999). A radiation hybrid breakpoint map of the acute myeloid leukemia (AML) and limb-girdle muscular

dystrophy 1A (LGMD1A) regions of chromosome 5q31 localizing 122 expressed sequences. Genomics **57**: 24-35.

Hosoe, S. (1996). Search for the tumor-suppressor gene(s) on chromosome 5q, which may play an important role for the progression of lung cancer. Nippon Rinsho **54**: 482-6.

Huebner, K., P. C. Nowell and C. M. Croce (1989). Lineage-specific gene rearrangement/deletion: a nonconservative model. Cancer Res **49**: 4071-4.

Itoh, S., H. Harada, Y. Nakamura, R. White and T. Taniguchi (1991). Assignment of the human interferon regulatory factor-1 (IRF1) gene to chromosome 5q23-q31. Genomics **10**: 1097-9.

Jaju, R. J., J. Boultwood, F. J. Oliver, M. Kostrzewa, C. Fidler, N. Parker, J. D. McPherson, S. W. Morris, U. Muller, J. S. Wainscoat and L. Kearney (1998). Molecular cytogenetic delineation of the critical deleted region in the 5q- syndrome. Genes Chromosomes Cancer **22**: 251-6.

Jaju, R. J., O. A. Haas, M. Neat, J. Harbott, V. Saha, J. Boultwood, J. M. Brown, H. Pirc-Danoewinata, B. W. Krings, U. Muller, S. W. Morris, J. S. Wainscoat and L. Kearney (1999). A new recurrent translocation, t(5;11)(q35;p15.5), associated with del(5q) in childhood acute myeloid leukemia. The UK Cancer Cytogenetics Group (UKCCG). Blood **94**: 773-80.

Jensen, S. J., E. P. Sulman, J. M. Maris, T. C. Matise, P. J. Vojta, J. C. Barrett, G. M. Brodeur and P. S. White (1997). An integrated transcript map of human chromosome 1p35-p36. Genomics **42**: 126-36.

Johansson, B., F. Mertens and F. Mitelman (1993). Cytogenetic deletion maps of hematologic neoplasms: circumstantial evidence for tumor suppressor loci. Genes Chromosomes Cancer **8**: 205-18.

Kaneko, H., S. Misawa, S. Horiike, H. Nakai and K. Kashima (1995). TP53 mutations emerge at early phase of myelodysplastic syndrome and are associated with complex chromosomal abnormalities. Blood **85**: 2189-93.

Karki, S., M. K. Tokito and E. L. Holzbaur (2000). A dynactin subunit with a highly conserved cysteine-rich motif interacts directly with Arp1. J Biol Chem **275**: 4834-9.

Kataoka, T. R., A. Ito, H. Asada, K. Watabe, K. Nishiyama, K. Nakamoto, S. Itami, K. Yoshikawa, M. Ito, H. Nojima and Y. Kitamura (2000). Annexin VII as a novel marker for invasive phenotype of malignant melanoma. Jpn J Cancer Res **91**: 75-83.

Kelm, R. J., Jr., G. A. Hair, K. G. Mann and B. W. Grant (1992). Characterization of human osteoblast and megakaryocyte-derived osteonectin (SPARC). Blood **80**: 3112-9.

Kennedy, D., T. Ramsdale, J. Mattick and M. Little (1996). An RNA recognition motif in Wilms' tumour protein (WT1) revealed by structural modelling. Nat Genet **12**: 329-31.

Kere, J., T. Ruutu and A. de la Chapelle (1987). Monosomy 7 in granulocytes and monocytes in myelodysplastic syndrome. N Engl J Med **316**: 499-503.

Klomp, L. W., S. J. Lin, D. S. Yuan, R. D. Klausner, V. C. Culotta and J. D. Gitlin (1997). Identification and functional expression of HAH1, a novel human gene involved in copper homeostasis. J Biol Chem **272**: 9221-6.

Knudson, A. G., Jr. (1971). Mutation and cancer: statistical study of retinoblastoma. Proc Natl Acad Sci U S A **68**: 820-3.

Konig, E. A., W. C. Kusser, C. Day, F. Porzsolt, B. W. Glickman, G. Messer, M. Schmid, R. de Chatel, Z. L. Marcsek and J. Demeter (2000). p53 mutations in hairy cell leukemia. Leukemia **14**: 706-11.

Kostrzewa, M., B. W. Krings, M. J. Dixon, K. Eppelt, A. Kohler, D. L. Grady, D. Steinberger, N. D. Fairweather, R. K. Moyzis, A. P. Monaco and U. Muller (1998). Integrated physical and transcript map of 5q31.3-qter. Eur J Hum Genet **6**: 266-74.

Kouides, P. A. and J. M. Bennett (1992). Morphology and classification of myelodysplastic syndromes. Hematol Oncol Clin North Am **6**: 485-99.

Krizman, D. B. and S. M. Berget (1993). Efficient selection of 3'-terminal exons from vertebrate DNA. Nucleic Acids Res **21**: 5198-202.

Kulkarni, S., C. Heath, S. Parker, A. Chase, S. Iqbal, C. F. Pocock, J. Kaeda, K. Cwynarski, J. M. Goldman and N. C. Cross (2000). Fusion of H4/D10S170 to the platelet-derived growth factor receptor beta in BCR-ABL-negative myeloproliferative disorders with a t(5;10)(q33;q21). Cancer Res **60**: 3592-8.

Kumar, T. R., L. A. Donehower, A. Bradley and M. M. Matzuk (1995). Transgenic mouse models for tumour-suppressor genes. J Intern Med **238**: 233-8.

Kurokawa, M., T. Tanaka, K. Tanaka, S. Ogawa, K. Mitani, Y. Yazaki and H. Hirai (1996). Overexpression of the AML1 proto-oncoprotein in NIH3T3 cells leads to neoplastic transformation depending on the DNA-binding and transactivational potencies. Oncogene **12**: 883-92.

La Spada, A. R., H. L. Paulson and K. H. Fischbeck (1994). Trinucleotide repeat expansion in neurological disease. Ann Neurol **36**: 814-22.

Lai, J. L., C. Preudhomme, M. Zandecki, M. Flactif, M. Vanrumbeke, P. Lepelley, E. Wattel and P. Fenaux (1995). Myelodysplastic syndromes and acute myeloid leukemia with 17p deletion. An entity characterized by specific dysgranulopoiesis and a high incidence of P53 mutations. Leukemia **9**: 370-81.

Lane, D. P., C. Midgley and T. Hupp (1993). Tumour suppressor genes and molecular chaperones. Philos Trans R Soc Lond B Biol Sci **339**: 369-72; discussion 372-3.

Lane, T. F. and E. H. Sage (1994). The biology of SPARC, a protein that modulates cell-matrix interactions. Faseb J **8**: 163-73.

Le Beau, M. M., R. Espinosa, 3rd, W. L. Neuman, W. Stock, D. Roulston, R. A. Larson, M. Keinanen and C. A. Westbrook (1993). Cytogenetic and molecular delineation of the smallest commonly deleted region of chromosome 5 in malignant myeloid diseases. Proc Natl Acad Sci U S A **90**: 5484-8.

Le Beau, M. M. (1992). Deletions of chromosome 5 in malignant myeloid disorders. Cancer Surv **15**: 143-59.

Le Beau, M. M., R. S. Lemons, R. Espinosa, 3rd, R. A. Larson, N. Arai and J. D. Rowley (1989). Interleukin-4 and interleukin-5 map to human chromosome 5 in a region encoding growth factors and receptors and are deleted in myeloid leukemias with a del(5q). Blood **73**: 647-50.

Ledda, M. F., S. Adris, A. I. Bravo, C. Kairiyama, L. Bover, Y. Chernajovsky, J. Mordoh and O. L. Podhajcer (1997). Suppression of SPARC expression by antisense RNA abrogates the tumorigenicity of human melanoma cells. Nat Med **3**: 171-6.

Legare, R. D. and D. G. Gilliland (1995). Myelodysplastic syndrome. Curr Opin Hematol **2**: 283-92.

Li, X., C. A. Wise, D. Le Paslier, A. L. Hawkins, C. A. Griffin, S. J. Pittler, M. Lovett and E. W. Jabs (1994). A YAC contig of approximately 3 Mb from human chromosome 5q31-->q33. Genomics **19**: 470-7.

Lin, S. J. and V. C. Culotta (1995). The ATX1 gene of Saccharomyces cerevisiae encodes a small metal homeostasis factor that protects cells against reactive oxygen toxicity. Proc Natl Acad Sci U S A **92**: 3784-8.

Liu, J., C. Wu, K. Bossie, K. Bejaoui, B. A. Hosler, J. C. Gingrich, M. Ben Hamida, F. Hentati, E. Schurr, P. J. de Jong and R. H. Brown, Jr. (1998). Generation of a 3-Mb PAC contig spanning the Miyoshi myopathy/limb- girdle muscular dystrophy (MM/LGMD2B) locus on chromosome 2p13. Genomics **49**: 23-9.

Liu, L., G. Shao, G. Steele-Perkins and S. Huang (1997). The retinoblastoma interacting zinc finger gene RIZ produces a PR domain-lacking product through an internal promoter. J Biol Chem **272**: 2984-91.

Luria, D., S. Avigad, I. J. Cohen, B. Stark, R. Weitz and R. Zaizov (1997). p53 mutation as the second event in juvenile chronic myelogenous leukemia in a patient with neurofibromatosis type 1. Cancer **80**: 2013-8.

Macera, M. J., C. J. Godec, N. Sharma and R. S. Verma (1999). Loss of heterozygosity of the TP53 tumor suppressor gene and detection of point mutations by the non-isotopic RNAse cleavage assay in prostate cancer. Cancer Genet Cytogenet **108**: 42-7.

Mahoney, P. A., U. Weber, P. Onofrechuk, H. Biessmann, P. J. Bryant and C. S. Goodman (1991). The fat tumor suppressor gene in Drosophila encodes a novel member of the cadherin gene superfamily. Cell **67**: 853-68.

Malone, K., M. M. Sohocki, L. S. Sullivan and S. P. Daiger (1999). Identifying and mapping novel retinal-expressed ESTs from humans. Mol Vis **5**: 5.

Mancini, D. N., S. M. Singh, T. K. Archer and D. I. Rodenhiser (1999). Site-specific DNA methylation in the neurofibromatosis (NF1) promoter interferes with binding of CREB and SP1 transcription factors. Oncogene **18**: 4108-19.

Mandla, S. G., S. Goobie, R. T. Kumar, O. Hayne, E. Zayed, D. L. Guernsey and W. L. Greer (1998). Genetic analysis of familial myelodysplastic syndrome: absence of linkage to chromosomes 5q31 and 7q22. Cancer Genet Cytogenet **105**: 113-8.

Maris J. M., J. Jensen, E. P. Sulman, C. P. Beltinger, C. Allen, J. A. Biegel, G. M. Brodeur, P. S. White (1997). Human Kruppel-related 3 (HKR3): a candidate for the 1p36 neuroblastoma tumour suppressor gene? Eur J Cancer **33**: 1991-6.

Mathew, P., A. Tefferi, G. W. Dewald, S. L. Goldberg, J. Su, H. C. Hoagland and P. Noel (1993). The 5q- syndrome: a single-institution study of 43 consecutive patients. Blood **81**: 1040-5.

Mayer, K., W. Ballhausen and H. D. Rott (1999). Mutation screening of the entire coding regions of the TSC1 and the TSC2 gene with the protein truncation test (PTT) identifies frequent splicing defects. Hum Mutat **14**: 401-11.

Melki, J. R., P. C. Vincent and S. J. Clark (1999). Cancer-specific region of hypermethylation identified within the HIC1 putative tumour suppressor gene in acute myeloid leukaemia. Leukemia **13**: 877-83.

Miyazato, A., S. Ueno, K. Ohmine, M. Ueda, K. Yoshida, Y. Yamashita, T. Kaneko, M. Mori, K. Kirito, M. Toshima, Y. Nakamura, K. Saito, Y. Kano, S. Furusawa, K. Ozawa and H. Mano (2001). Identification of myelodysplastic syndrome-specific genes by DNA microarray analysis with purified hematopoietic stem cell fraction. Blood **98**: 422-7.

Miyoshi, H., M. Ohira, K. Shimizu, K. Mitani, H. Hirai, T. Imai, K. Yokoyama, E. Soeda and M. Ohki (1995). Alternative splicing and genomic structure of the AML1 gene involved in acute myeloid leukemia. Nucleic Acids Res **23**: 2762-9.

Mok, S. C., W. Y. Chan, K. K. Wong, M. G. Muto and R. S. Berkowitz (1996). SPARC, an extracellular matrix protein with tumor-suppressing activity in human ovarian epithelial cells. Oncogene **12**: 1895-901.

Mok, S. C., K. W. Lo and S. W. Tsao (1993). Direct cycle sequencing of mutated alleles detected by PCR single- strand conformation polymorphism analysis. Biotechniques **14**: 790-4.

Morgan, S. E. and M. B. Kastan (1997). p53 and ATM: cell cycle, cell death, and cancer. Adv Cancer Res **71**: 1-25.

Mufti, G. J. (1992). Chromosomal deletions in the myelodysplastic syndrome. Leuk Res **16**: 35-41.

Mundel, P., H. W. Heid, T. M. Mundel, M. Kruger, J. Reiser and W. Kriz (1997). Synaptopodin: an actin-associated protein in telencephalic dendrites and renal podocytes. J Cell Biol **139**: 193-204.

Murthy, A. E., A. Bernards, D. Church, J. Wasmuth and J. F. Gusella (1996). Identification and characterization of two novel tetratricopeptide repeat-containing genes. DNA Cell Biol **15**: 727-35.

Nagarajan, L., J. Zavadil, D. Claxton, X. Lu, J. Fairman, J. A. Warrington, J. J. Wasmuth, A. C. Chinault, C. E. Sever, M. L. Slovak and et al. (1994). Consistent loss of the D5S89 locus mapping telomeric to the interleukin gene cluster and centromeric to EGR-1 in patients with 5q- chromosome. Blood **83**: 199-208.

Nakayama, M., D. Nakajima, T. Nagase, N. Nomura, N. Seki and O. Ohara (1998). Identification of high-molecular-weight proteins with multiple EGF-like motifs by motif-trap screening. Genomics **51**: 27-34.

Neubauer, A., C. Brendel, D. Vogel, C. A. Schmidt, I. Heide and D. Huhn (1993). Detection of p53 mutations using nonradioactive SSCP analysis: p53 is not frequently mutated in myelodysplastic syndromes (MDS). Ann Hematol **67**: 223-6.

Nutt, S. L. and M. Busslinger (1999). Monoallelic expression of Pax5: a paradigm for the haploinsufficiency of mammalian Pax genes? Biol Chem **380**: 601-11.

O'Donovan, M. C., P. J. Oefner, S. C. Roberts, J. Austin, B. Hoogendoorn, C. Guy, G. Speight, M. Upadhyaya, S. S. Sommer and P. McGuffin (1998). Blind analysis of denaturing high-performance liquid chromatography as a tool for mutation detection. Genomics **52**: 44-9.

Orlow, I., L. Lacombe, G. J. Hannon, M. Serrano, I. Pellicer, G. Dalbagni, V. E. Reuter, Z. F. Zhang, D. Beach and C. Cordon-Cardo (1995). Deletion of the p16 and p15 genes in human bladder tumors. J Natl Cancer Inst **87**: 1524-9.

Pappas, G. J., M. H. Polymeropoulos, J. M. Boyle and J. M. Trent (1995). Regional assignment by hybrid mapping of 36 expressed sequence tags (ESTs) on human chromosome 6. Genomics **25**: 124-9.

Parsons, R. (1998). Phosphatases and tumorigenesis. Curr Opin Oncol **10**: 88-91.

Passmore, S. J., I. M. Hann, C. A. Stiller, P. Ramani, G. J. Swansbury, B. Gibbons, B. R. Reeves and J. M. Chessells (1995). Pediatric myelodysplasia: a study of 68 children and a new prognostic scoring system. Blood **85**: 1742-50.

Pearson, P. L. and R. B. Van der Luijt (1998). The genetic analysis of cancer. J Intern Med **243**: 413-7.

Pearson, W. R. and D. J. Lipman (1988). Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A **85**: 2444-8.

Pedersen, B. and I. M. Jensen (1991). Clinical and prognostic implications of chromosome 5q deletions: 96 high resolution studied patients. Leukemia **5**: 566-73.

Permana, P. A. and D. M. Mott (1997). Genetic analysis of human type 1 protein phosphatase inhibitor 2 in insulin-resistant Pima Indians. Genomics **41**: 110-4.

Philipp-Staheli, J., S. R. Payne and C. J. Kemp (2001). p27(Kip1): regulation and function of a haploinsufficient tumor suppressor and its misregulation in cancer. Exp Cell Res **264**: 148-68.

Pritchard-Jones, K. and L. King-Underwood (1997). The Wilms tumour gene WT1 in leukaemia. Leuk Lymphoma **27**: 207-20.

Rand, J. H. (1999). "Annexinopathies"--a new class of diseases. N Engl J Med **340**: 1035-6.

Roberts, P. S., S. Jozwiak, D. J. Kwiatkowski and S. L. Dabora (2001). Denaturing high-performance liquid chromatography (DHPLC) is a highly sensitive, semi-automated method for identifying mutations in the TSC1 gene. J Biochem Biophys Methods **47**: 33-7.

Robertson, K. D., S. Ait-Si-Ali, T. Yokochi, P. A. Wade, P. L. Jones and A. P. Wolffe (2000). DNMT1 forms a complex with Rb, E2F1 and HDAC1 and represses transcription from E2F-responsive promoters. Nat Genet **25**: 338-42.

Rotig, A., I. Valnot, C. Mugnier, P. Rustin and A. Munnich (2000). Screening human EST database for identification of candidate genes in respiratory chain deficiency. Mol Genet Metab **69**: 223-32.

Rubin, G. M., M. D. Yandell, J. R. Wortman, G. L. Gabor Miklos, C. R. Nelson, I. K. Hariharan, M. E. Fortini, P. W. Li, R. Apweiler, W. Fleischmann, J. M. Cherry, S. Henikoff, M. P. Skupski, S. Misra, M. Ashburner, E. Birney, M. S. Boguski, T. Brody, P. Brokstein, S. E. Celniker, S. A. Chervitz, D. Coates, A. Cravchik, A. Gabrielian, R. F. Galle, W. M. Gelbart, R. A. George, L. S. Goldstein, F. Gong, P. Guan, N. L. Harris, B. A. Hay, R. A. Hoskins, J. Li, Z. Li, R. O. Hynes, S. J. Jones, P. M. Kuehl, B. Lemaitre, J. T. Littleton, D. K. Morrison, C. Mungall, P. H. O'Farrell, O. K. Pickeral, C. Shue, L. B. Vosshall, J. Zhang, Q. Zhao, X. H. Zheng and S. Lewis (2000). Comparative genomics of the eukaryotes. Science **287**: 2204-15.

Sage, E. H. and P. Bornstein (1991). Extracellular proteins that modulate cell-matrix interactions. SPARC, tenascin, and thrombospondin. J Biol Chem **266**: 14831-4.

Saiki, R. K., D. H. Gelfand, S. Stoffel, S. J. Scharf, R. Higuchi, G. T. Horn, K. B. Mullis and H. A. Erlich (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science **239**: 487-91.

Saitou, M., J. Sugimoto, T. Hatakeyama, G. Russo and M. Isobe (2000). Identification of the TCL6 genes within the breakpoint cluster region on chromosome 14q32 in T-cell leukemia. Oncogene **19**: 2796-802.

Sampson, J. R. and P. C. Harris (1994). The molecular genetics of tuberous sclerosis. Hum Mol Genet **3**: 1477-80.

Sanger, F., S. Niklen and A. R. Coulson (1977). DNA sequencing with chain terminating inhibitors. Proc Natl Acad Sci USA **74**: 5463-5467.

Sargent, C. A., N. A. Affara, E. Bentley, A. Pelmear, D. M. Bailey, P. Davey, D. Dow, M. Leversha, H. Aplin, G. T. Besley and et al. (1993). Cloning of the X-linked glycerol kinase deficiency gene and its identification by sequence comparison to the Bacillus subtilis homologue. Hum Mol Genet **2**: 97-106.

Sato, K., G. Tamura, T. Tsuchiya, Y. Endoh, O. Usuba, W. Kimura and T. Motoyama (2001). Frequent loss of expression without sequence mutations of the DCC gene in primary gastric cancer. Br J Cancer **85**: 199-203.

Schwaller, J., T. Pabst, H. P. Koeffler, G. Niklaus, P. Loetscher, M. F. Fey and A. Tobler (1997). Expression and regulation of G1 cell-cycle inhibitors (p16INK4A, p15INK4B, p18INK4C, p19INK4D) in human acute myeloid leukemia and normal myeloid cells. Leukemia **11**: 54-63.

Sharan, S. K. and A. Bradley (1998). Functional characterization of BRCA1 and BRCA2: clues from their interacting proteins. J Mammary Gland Biol Neoplasia **3**: 413-21.

Sherr, C. J., C. W. Rettenmier, R. Sacca, M. F. Roussel, A. T. Look and E. R. Stanley (1985). The c-fms proto-oncogene product is related to the receptor for the mononuclear phagocyte growth factor, CSF-1. Cell **41**: 665-76.

Sill, H., R. C. Aguiar, H. Schmidt, A. Hochhaus, J. M. Goldman and N. C. Cross (1996). Mutational analysis of the p15 and p16 genes in acute leukaemias. Br J Haematol **92**: 681-3.

Simmons, A. D., S. A. Goodart, T. D. Gallardo, J. Overhauser and M. Lovett (1995). Five novel genes from the cri-du-chat critical region isolated by direct selection. Hum Mol Genet **4**: 295-302.

Skuse, G. R. and J. W. Ludlow (1995). Tumour suppressor genes in disease and therapy. Lancet **345**: 902-6.

Smith, P. D., A. Davies, M. J. Crumpton and S. E. Moss (1994). Structure of the human annexin VI gene. Proc Natl Acad Sci U S A **91**: 2713-7.

Snell, R. G., L. A. Doucette-Stamm, K. M. Gillespie, S. A. Taylor, L. Riba, G. P. Bates, M. R. Altherr, M. E. MacDonald, J. F. Gusella, J. J. Wasmuth and et al. (1993). The isolation of cDNAs within the Huntington disease region by hybridisation of yeast artificial chromosomes to a cDNA library. Hum Mol Genet **2**: 305-9.

Soenen, V., C. Preudhomme, C. Roumier, A. Daudignon, J. L. Lai and P. Fenaux (1998). 17p Deletion in acute myeloid leukemia and myelodysplastic syndrome. Analysis of breakpoints and deleted segments by fluorescence in situ. Blood **91**: 1008-15.

Soenen, V., C. Preudhomme, C. Roumier, J. L. Lai, P. Lepelley, T. Facon, D. Pagniez and P. Fenaux (1998). Myelodysplasia during the course of myeloma. Restriction of 17p deletion and p53 overexpression to myeloid cells. Leukemia **12**: 238-41.

Sokal, G., J. L. Michaux, H. Van Den Berghe, A. Cordier, J. Rodhain, A. Ferrant, M. Moriau, M. De Bruyere and J. Sonnet (1975). A new hematologic syndrome with a distinct karyotype: the 5 q-- chromosome. Blood **46**: 519-33.

Su, G., T. Roberts and J. K. Cowell (1999). TTC4, a novel human gene containing the tetratricopeptide repeat and mapping to the region of chromosome 1p31 that Is frequently deleted in sporadic breast cancer. Genomics **55**: 157-63.

Swaroop, A., B. L. Hogan and U. Francke (1988). Molecular analysis of the cDNA for human SPARC/osteonectin/BM-40: sequence, expression, and localization of the gene to chromosome 5q31- q33. Genomics **2**: 37-47.

Tannapfel, A., M. Benicke, A. Katalinic, D. Uhlmann, F. Kockerling, J. Hauss and C. Wittekind (2000). Frequency of p16(INK4A) alterations and K-ras mutations in intrahepatic cholangiocarcinoma of the liver. Gut **47**: 721-7.

The Huntington's Disease Collaborative Research Group (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell **72**: 971-83.

Theobald, J., A. Hanby, K. Patel and S. E. Moss (1995). Annexin VI has tumour-suppressor activity in human A431 squamous epithelial carcinoma cells. Br J Cancer **71**: 786-8.

Theobald, J., P. D. Smith, S. M. Jacob and S. E. Moss (1994). Expression of annexin VI in A431 carcinoma cells suppresses proliferation: a possible role for annexin VI in cell growth regulation. Biochim Biophys Acta **1223**: 383-90.

Tien, H. F., J. H. Tang, W. Tsay, M. C. Liu, F. Y. Lee, C. H. Wang, Y. C. Chen and M. C. Shen (2001). Methylation of the p15(INK4B) gene in myelodysplastic syndrome: it can be detected early at diagnosis or during disease progression and is highly associated with leukaemic transformation. Br J Haematol **112**: 148-54.

Todd, R., B. Bia, E. Johnson, C. Jones and F. Cotter (2001). Molecular characterization of a myelodysplasia-associated chromosome 7 inversion. Br J Haematol **113**: 143-52.

Trower, M. K., S. M. Orton, I. J. Purvis, P. Sanseau, J. Riley, C. Christodoulou, D. Burt, C. G. See, G. Elgar, R. Sherrington, E. I. Rogaev, P. St George-Hyslop, S. Brenner and C. W. Dykes (1996). Conservation of synteny between the genome of the pufferfish (Fugu rubripes) and the region on human chromosome 14 (14q24.3) associated with familial Alzheimer disease (AD3 locus). Proc Natl Acad Sci U S A **93**: 1366-9.

Tugendreich, S., M. S. Boguski, M. S. Seldin and P. Hieter (1993). Linking yeast genetics to mammalian genomes: identification and mapping of the human homolog of CDC27 via the expressed sequence tag (EST) data base. Proc Natl Acad Sci U S A **90**: 10031-5.

Van Cong, N., S. Fichelson, M. S. Gross, B. Sola, D. Bordereaux, M. F. de Tand, S. Guilhot, S. Gisselbrecht, J. Frezal and P. Tambourin (1989). The human homologues of Fim1, Fim2/c-Fms, and Fim3, three retroviral integration regions involved in mouse myeloblastic leukemias, are respectively located on chromosomes 6p23, 5q33, and 3q27. Hum Genet **81**: 257-63.

Van den Berghe, H., J. J. Cassiman, G. David, J. P. Fryns, J. L. Michaux and G. Sokal (1974). Distinct haematological disorder with deletion of long arm of no. 5 chromosome. Nature **251**: 437-8.

Van den Berghe, H., K. Vermaelen, C. Mecucci, D. Barbieri and G. Tricot (1985). The 5q-anomaly. Cancer Genet Cytogenet **17**: 189-255.

Van den Bosch, B. J., R. F. de Coo, H. R. Scholte, J. G. Nijland, R. van Den Bogaard, M. de Visser, C. E. de Die-Smulders and H. J. Smeets (2000). Mutation analysis of the entire mitochondrial genome using denaturing high performance liquid chromatography. Nucleic Acids Res **28**: E89.

Villarreal, X. C., K. G. Mann and G. L. Long (1989). Structure of human osteonectin based upon analysis of cDNA and genomic sequences. Biochemistry **28**: 6483-91.

Wallace, M. R., D. A. Marchuk, L. B. Andersen, R. Letcher, H. M. Odeh, A. M. Saulino, J. W. Fountain, A. Brereton, J. Nicholson, A. L. Mitchell and et al. (1990). Type 1 neurofibromatosis gene: identification of a large transcript disrupted in three NF1 patients. Science **249**: 181-6.

Wang-Gohrke, S., S. Hees, A. Pochon, W. H. Wen, A. Reles, M. F. Press, R. Kreienberg and I. B. Runnebaum (1998). Genomic semi-automated cycle sequencing as a sensitive screening technique for p53 mutations in frozen tumor samples. Oncol Rep **5**: 65-8.

Waters, J. C. and E. D. Salmon (1995). Chromosomes take an active role in spindle assembly. Bioessays **17**: 911-4.

Williams, R. T., S. S. Manji, N. J. Parker, M. S. Hancock, L. Van Stekelenburg, J. P. Eid, P. V. Senior, J. S. Kazenwadel, T. Shandala, R. Saint, P. J. Smith and M. A. Dziadek (2001). Identification and characterization of the STIM (stromal interaction molecule) gene family: coding for a novel class of transmembrane proteins. Biochem J **357**: 673-85.

Willis, T. G., I. R. Zalcberg, L. J. Coignet, I. Wlodarska, M. Stul, D. M. Jadayel, C. Bastard, J. G. Treleaven, D. Catovsky, M. L. Silva and M. J. Dyer (1998). Molecular cloning of translocation t(1;14)(q21;q32) defines a novel gene (BCL9) at chromosome 1q21. Blood **91**: 1873-81.

Willman, C. L., C. E. Sever, M. G. Pallavicini, H. Harada, N. Tanaka, M. L. Slovak, H. Yamamoto, K. Harada, T. C. Meeker, A. F. List and et al. (1993). Deletion of IRF-1, mapping to chromosome 5q31.1, in human leukemia and preleukemic myelodysplasia. Science **259**: 968-71.

Wimmer, K., M. Eckart, P. F. Stadler, H. Rehder and C. Fonatsch (2000). Three different premature stop codons lead to skipping of exon 7 in neurofibromatosis type I patients. Hum Mutat **16**: 90-1.

Woodcock, D. M., M. E. Linsenmeyer, J. P. Doherty and W. D. Warren (1999). DNA methylation in the promoter region of the p16 (CDKN2/MTS-1/INK4A) gene in human breast tumours. Br J Cancer **79**: 251-6.

Wu, Y., G. C. Fraizer and G. F. Saunders (1995). GATA-1 transactivates the WT1 hematopoietic specific enhancer. J Biol Chem **270**: 5944-9.

Yoshimura, S., H. Suemizu, Y. Taniguchi, K. Arimori, N. Kawabe and T. Moriuchi (1994). The human plasma glutathione peroxidase-encoding gene: organization, sequence and localization to chromosome 5q32. Gene **145**: 293-7.

Yuasa, Y. (2000). [Hereditary nonpolyposis colorectal cancer]. Nippon Rinsho **58**: 1396-9.

Zhang, Q. and K. Minoda (1995). Mutation detection and genetic counseling in retinoblastoma using heteroduplex analysis. Jpn J Ophthalmol **39**: 432-7.

Appendix

Stock solutions and Buffers

1. Separation of granulocytes and mononuclear cells by density gradient centrifugation.

a. Phosphate Buffered Saline (PBS)

PBS tablets (Sigma Aldrich)	5
PBS-EDTA stock	10mls
Distilled water up to 1 litre	
Autoclave	

b. PBS-EDTA stock

PBS tablets	5
0.5M EDTA (pH8.0)	10mls
Distilled water up to 1 litre	
Autoclave	

c. 0.5M EDTA (pH8.0) 1 litre

EDTA	186.1g

Dissolve in 800mls of distilled water using a magnetic plate and flea. Adjust pH to 8.0 with NaOH (~ 20g NaOH pellets). Make up to 1 litre with distilled water and autoclave.

d. Red cell lysis buffer (pH 7.2)

Sodium bicarbonate	1g
Ammonium chloride	8.29g
0.5M EDTA (pH 8.0)	200µl
Distilled water to 1 litre	

The solution was prepared freshly when required and filter sterilised prior to use (0.22µM filter, Falcon).

2. Standard restriction enzyme digestion of genomic DNA, and gel electrophoresis

a. 10x TBE (1 litre)

Trisma base	108g
Boric acid	55g
EDTA	9.3g
Distilled water to 1 litre	

b. Gel loading dye (50mls)

Bromophenol blue	0.25%
Xylene cyanol FF	0.25%
Ficoll (Type 400)	15%
Distilled water to 50mls	

3. Southern blotting

a. Ethidium bromide solution

1 ethidium bromide tablet (100mg) was dissolved in 10mls of distilled water in a fume cupboard. The solution was stored at room temperature protected from the light.

b. Denaturing solution

NaCl	1.5M
NaOH	0.5M

c. Alkali transfer buffer

NaCl	1.5M
NaOH	0.25M

d. Neutralisation solution

Tris-HCl (pH 7.5)	1M
NaCl	3M

4. Transformation of competent cells

a. LB (Luria Bertani) media (1 litre)

Bacto-tryptone	10g
Bacto yeast extract	5g
NaCl	10g

Distilled water to 1 litre

Autoclave

b. LB media with agar

Same as for LB media, but just before autoclaving add 15g/litre bacto-agar. Allow to cool to 50°C and add appropriate antibiotics if necessary. Pour plates immediately allowing approximately 30-35mls of medium per 90mm plate.

5. Recovery of the probe from the plasmid

a. 50x TAE (1 litre)

Trisma base	242g
Glacial acetic acid	57.1mls
0.5M EDTA (pH 8.0)	100mls

Distilled water to 1 litre

6. Probe labelling

a. Sephadex grade G-100

Sephadex grade G-100 powder was added to sterile distilled water. 10g of powder yielded approximately 160ml of slurry. The swollen resin was washed several times with sterile distilled water to remove soluble dextran (the volume of water used was at least twice the volume of resin). The resin was finally equilibrated in TE buffer (pH 8) and stored at room temperature.

b. TE buffer (pH 8.0)

10mM Tris pH 8.0

1mM EDTA pH 8.0

c. 1M Tris (1 litre)

Trisma base 121.1g

pH adjusted to the desired value by the addition of concentrated HCl.

Autoclave

7. Filter hybridisation

a. 20x SSC

NaCl	3M
Tri-sodium citrate	0.3M

b. Hybridisation buffer

SSPE	5x
SDS	1%
Denhardt's solution	5x
Dextran sulphate	5%

Aliquot into 50ml centrifuge tubes and store at -20°C. Prior to use, defrost appropriate volume and add 100mg/ml denatured salmon testes DNA (Sigma Aldrich).

c. SSPE (20x stock solution) 1 litre

NaCl	175.3g
$NaH_2PO_4.H_2O$	27.6g
EDTA	7.4g

Dissolve in 800mls of distilled water. Adjust the pH to 7.4 with NaOH and make up to 1 litre with distilled water and autoclave.

d. Denhardt's reagent (50x stock solution)

Ficoll (Type 400)	5g
Polyvinylpyrollidone	5g
Bovine serum albumin (BSA)	5g

Distilled water to 500mls. Filter sterilise with a 0.22µ filter, Falcon, and aliquot into 50ml centrifuge tubes. Store at -20°C.

8. Preparation of competent *E.coli* for transformation

a. M9 minimal media (500mls)

Na_2HPO_4	6g
KH_2PO_4	3g
NH_4Cl	1g
NaCl	0.5g

Distilled water to 500mls

Autoclave and cool to 55°C

1M $MgSO_4$	1ml
Glucose	2g
1M $CaCl_2$	0.1ml
Thiamine	0.34g

Adjust volume to 10mls with distilled water and filter sterilise. Add to the cooled M9 media.

b. Minimal media agar plates

As above but with 8g of bacto-agar. Prepare M9 media in a total volume of 300mls of distilled water and dissolve 8g bacto-agar in the remaining 200mls. Autoclave separately. Mix the solutions together after autoclaving and cool to 55°C. Add 10mls of filter sterilised solution as above and pour plates.

c. SOB media (1 litre)

Bacto-tryptone	20g
Bacto yeast extract	5g
NaCl	0.5g

Distilled water to 1 litre

Autoclave and add 20mls of sterile 1M $MgSO_4$.

d. SOB media with agar

As above but just prior to autoclaving add 15g/litre of bacto-agar. Cool to 50°C before pouring plates.

e. TFB (1 litre)

Equilibrate a 0.5M solution of MES (2(N-morpholino)ethone sulphonic acid) to pH 6.3 using KOH pellets and sterilise by filtration. Store in aliquots at -20°C.

0.5M K-MES pH 6.3	20mls
KCl (ultrapure)	7.4g
$MnCl_2.4H_2O$	8.9g
$CaCl_2.2H_2O$	1.5g
Hexa-amine cobalt chloride	0.8g

Make up the solution using the purest available water and add the salts as solids. Sterilise by filtration into 50ml aliquots and store at 4°C.

9. Transformation of competent cells with ligated M13

a. 2x YT media (1 litre)

Bacto-tryptone	16g
Bacto yeast extract	10g
NaCl	5g

Distilled water to 950mls. Adjust pH to 7.0 with NaOH. Distilled water to 1 litre and autoclave.

b. 2 x YT agar

Same as for 2x YT media, but prior to autoclaving, add 15g/litre bacto-agar. Cool to 50°C before pouring plates.

c. 2 x YT soft agar

Same as for 2 X YT media, but prior to autoclaving add 7g/litre bacto-agar.

d. X-gal

5-Bromo-4-chloro-3-indolyl-ß-D-galactoside.

Dissolve X-gal in dimethylformamide to make a 2% solution. Cover tube in foil to protect from the light and store at –20°C.

e. IPTG

Isopropylthio-ß-D-galactoside. Dissolve IPTG in distilled water to make a 2% solution. Sterilise by filtration through a 0.22µ disposable filter. Aliquot and store at –20°C.

10. Preparation of single-stranded templates

a. 3M Sodium acetate

Sodium acetate.$3H_2O$ 408.1g

Dissolve in 800mls of distilled water. Adjust the pH to 5.2 with glacial acetic acid or to 7.0 with dilute acetic acid. Make volume up to 1 litre with distilled water and autoclave.

11. Preparation of the sequencing gel plates

a. Bind silane

Absolute ethanol	2mls
10% acetic acid	0.5mls
Bind silane	7.5µl

Prepare immediately before use

12. Transformation of competent cells with ligated pGEM®-T Easy

a. 2mM Mg^{2+} stock

$MgCl_2 \bullet 6H_2O$	20.33g
$MgSO_4 \bullet 7H_2O$	224.65g

Add distilled water to 100ml. Filter sterilise

b. SOC media

Bacto®-tryptone	2.0g
Bacto®-yeast extract	0.5g
1M Nacl	1ml
1M KCl	0.25ml
2M Mg2+ stock, filter-sterilised	1ml
2M glucose, filter sterilised	1ml

Add Bacto®-tryptone, Bacto®-yeast extract, NaCl and KCl to 97ml distilled water. Stir to dissolve. Autoclave and cool to room temperature. Add 2M Mg2+ stock and 2M glucose, each to a final concentration of 20mM. Bring to 100ml with sterile, distilled water. Filter the complete medium through a 0.2µm filter unit. The final pH should be 7.0.

13. Precipitation of sequencing reactions with the Thermo Sequenase™ Cy™5 Dye Terminator Kit (Amersham Pharmacia Biotech)

a. 7.5M Ammonium acetate (50mls)

Ammonium acetate	28.905g

Distilled water to 50mls

14. The WAVE™ DNA fragment analysis system

a. Buffer A

0.1M Triethylammonium acetate (TEAA)	50mls
Acetonitrile (HPLC Grade)	250µl
Milli-Q water to 1 litre	

b. Buffer B

0.1M Triethylammonium acetate (TEAA)	50mls
25% Acetonitrile (HPLC Grade)	250mls
Milli-Q water to 950mls	

Mix by inversion. Solution undergoes an endothermic reaction when TEAA mixes with Acetonitrile. When solution has warmed to room temperature, add Milli-Q water to 1 litre.

c. Buffer C (75% Acetonitrile wash solution)

75% Acetonitrile (HPLC Grade)	750mls
Milli-Q water to 1 litre	

d. Buffer D (8% Acetonitrile syringe wash solution)

8% Acetonitrile (HPLC Grade)	80mls
Milli-Q water to 1 litre	